Early Prediction of Sepsis using Time Series Forecasting

Jinghua Xu*§, Natalia Minakova†§, Pablo Ortega Sanchez‡§ and Stefan Riezler¶
Department of Computational Linguistics
Heidelberg University
Heidelberg, Germany
Email: *xu@cl.uni-heidelberg.de, †mina@cl.uni-heidelberg.de, ‡sanchez@cl.uni-heidelberg.de, ¶riezler@cl.uni-heidelberg.de

2023 IEEE 19th International Conference on e-Science (e-Science) | 979-8-3503-2223-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/E-SCIENCE58273.2023.10254852

***Abstract*—Sepsis is a serious complication of an infection. Without quick treatment it can lead to organ failure and death. Early detection and treatment of sepsis can thus improve patient outcomes. Yet, their effectiveness often relies on awareness and acceptance of said procedures. In this work, we implement sepsis check based on a widely accepted guideline for sepsis recognition (Sepsis-3). Our implementation achieved F-score as high as 0.874. In addition to implementing the ruled-based approach to early sepsis detection, we use an existing data-driven transformer-based STraTS model [1] for time-series forecasting to support sepsis check and directly predicting sepsis label using 24-hour patient data in a fully data-driven setup. The advantage of time series forecasting is improved handling of missing data and the potential of applying the Sepsis-3 definition to unobserved forecast data. Additionally, we attempt to improve STraTS model by integrating a clinical text embedding module to enable multimodal learning. Both the original STraTS model and our refined STraTS+Text model perform good in both forecasting (masked MSE, mean squared error at approximately 5.24) and classification task (ROC-AUC, area under receiver operating characteristic curve at approximately 0.89).**

***Index Terms*—deep learning, transformer, sepsis**

I. Introduction

Sepsis occurs when the body's immune response to an infection becomes dysregulated and triggers systemic inflammation. It is a leading cause of death in the Intensive Care Units (ICU). Early detection of sepsis is crucial for patient survival [2]. In an effort to identify septic patients from their clinical data, [3] and [4] present slightly different approaches. These rule-based guidelines revolve around identifying suspected infections and a clinical criterion for life-threatening organ dysfunction. The guidelines by [3] were developed as an in-hospital tool to determine the state and condition of a patient. The implementation of these guidelines can be applied to already observed data, but more importantly, could be applied to forecast time-series values based on observed data. This would potentially allow us to identify the development of sepsis earlier and maybe even improve the chances of preventing sepsis. Therefore, we implement a rule-based sepsis check based on a widely accepted guideline for sepsis recognition (Sepsis-3) to enable early sepsis prediction.

In addition to implementing a rule-based sepsis check, we use and refine an existing Self-supervised Transformer for Time-Series (STraTS) model [1] for time-series forecasting and 24-hour sepsis prediction. We use the STraTS regression model to forecast time-series values following each observation window to support sepsis check, and the STraTS classification model to predict sepsis label using 24-hour patient data in a fully data-driven setup. On the basis of the original STraTS architecture, which can only take continuous physiological features as its input, we integrate a clinical text embedding module based on Clinical BERT [5] to encode 1.4 million clinical notes associated with patients in our data from MIMIC-III. Both models (STraTS and STraTS+Text) show good performance in both forecasting and classification tasks, achieving masked MSE (Mean Squared Error) approximately at 5.24 and ROC-AUC (area under receiver operating characteristic curve) approximately at 0.89.

Our rule-based sepsis check achieved an F-score as high as 0.874 without using features values predicted by the STraTS forecasting model. While introducing predicted values produced by the STraTS forecasting model did not improve the performance of the rule-based sepsis check, the STraTS forecasting predictions helped with the problem of data sparsity and enabled the rule-based sepsis check to identify septic patients whose clinical data alone would not be sufficient to correctly classify them.

We release our code at https://github.com/JINHXu/Early-Sepsis-Prediction-using-TSF.

II. Related Work

In [6], it was found that in 2011, sepsis was the most expensive condition treated in U.S. hospitals and accounted for 5.2% of the total hospitalization costs. In a more global context, [2] found that in 2017 sepsis resulted in 19.7% of all global deaths, with the most cases in low- or middle income countries. While guidelines for identifying and treating septic patients do exist, like the one proposed by [3], [7] argue that many factors complicate their on-site implementation. Depending on the environment and resources, these factors include critical staff shortages, failure to identify sepsis and lack of availability of laboratory tests. Previous hypothesis regarding the development and diagnosis of sepsis heavily relied on the Systemic Inflammatory Response-Syndrome (SIRS) [8]. More recently, the reliability of SIRS has been found to be not

§Equal contribution. Authors listed alphabetically.

sufficient which prompted the development of the Sepsis-related Organ Failure Assessment (SOFA) [9].

With the growing use of electronic health record (EHR) systems, clinical time-series data and digital clinical notes have been used to enable machine learning models for prediction tasks in the medical domain such as sepsis [10] and mortality prediction [1]. Earlier studies applied linear dynamical systems (LDS) [11] and Gaussian process (GP) [12] to model time-series data in the clinical domain. Meanwhile models such as BioBERT [13] and Clinical BERT [5] pretrained on text data in the medical domain have been proposed to model articles and doctor-written notes in the medical domain.

Moving past the phase of simply modeling data, more recent studies seek to solve the problem in specific predictive tasks. In [10], it proposed a multi-modal learning solution to early sepsis prediction by integrating clinical notes and continuous physiological features to construct a transformer-based binary classification model. The deep learning architecture employs a PTSM (Physiological Time Series Model) module to encode physiological features over time, and a Clinical BERT [5] text embedding module to represent clinical notes. While the model in [10] achieved good performance on MIMIC-III [14] and eICU-CRD [15] datasets to predict a binary label indicating patient septic state based on observed data following admission into ICU, it functions simply as a binary classification model and is not capable of time-series forecasting. Additionally, the approach proposed in [10] simply deals with missing values in time-series data through imputation by forward and backward filling. Whereas, in [1], it proposed a two-step transformer-based approach to mortality prediction, including a regression model to forecast time-series values and a classification model to predict patient mortality. Compared to the model proposed in [10], STraTS [1] is not only capable of time series forecasting, which allows it to assist a Sepsis-3-based check, but also able to dealing with sporadic clinical data through a novel triplet embedding module to avoid data imputation commonly used in other methods such as in [10]. Thus in this work, we follow STraTS over other existing methods for both time-series forecasting to support our rule-based sepsis check, and directly use the classification model for sepsis prediction. However, it is worth noting that the original STraTS model represents a single-modal learning paradigm, i.e. it considers only physiological features and disregards information contained in clinical notes. Thus, in this work, we attempt to integrate a text embedding module to encode clinical notes and enable multi-modal learning.

III. Sepsis-3 Implementation

A. Suspected Infection

According to the third International Consensus Definitions for Sepsis and Septic Shock [3], Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection. Thus, one of the key indicators for septic patients is a suspected infection. With only data at hand, a suspected infection is identified by orders for blood cultures and antibiotics. The guideline by [3] only accounts for a specific time period in which antibiotics and cultures are ordered. If antibiotics were administered first, blood cultures need to be taken within 24 hours and if blood cultures were taken first, antibiotics need to be administered within 72 hours. In the implementation, this time period is called suspicion window. In contrast, [4] require antibiotics to be administered for at least 72 consecutive hours to be considered as a suspected infection. If that is the case, the first administration of antibiotics is compared to blood culture orders identically to [3]. This distinction has a rather great impact on the implementation, as the sparsity of available data and therefore the ability to identify consecutive administrations of antibiotics is a non-trivial challenge.

B. Criterion for life-threatening organ dysfunction

Another key indicator for septic patients is life-threatening organ dysfunction. There are several criterions that can be employed to identify life-threatening organ dysfunction, however, both [3] and [4] suggest the Sequential [Sepsis-related] Organ Function Assessment (SOFA) [9], which takes into account a variety of clinical and laboratory variables. Following [3], a patient's SOFA score should be computed for each time step – which in this case means per hour. The time of SOFA is then the time at which a patient gets a SOFA score of two or higher, considering the initial value to be zero if no organ dysfunction is known beforehand. [4] consider the time of SOFA to be the time at which an increase of two in comparison to the last 24 hours occurs. If this time of SOFA is at most 48 hours before or 24 hours after a suspected infection, the patient developed sepsis according to [3]. [4] are more strict and only allow the time of SOFA to be at most 24 hours before and 12 hours after the time of a suspected infection. Additionally, they treat the earlier of the two times as the time of onset of sepsis. The aforementioned time period between suspected infection and time of SOFA is called sepsis window in the implementation.

C. Implementation

In order to compute checks for either [3] or [4], the clinical features described in Tables X and XI should be reported in the patient data, with a great importance on antibiotics and blood culture features. Given the nature of the rule-based guidelines, the lack of either the antibiotics or blood culture feature will always result in a negative sepsis label, as these features are directly responsible for suspecting an infection.

D. About preprocessing and running the sepsis check

The sepsis check comprises several utility functions to process [1]'s output. Before starting an experiment one needs to decide if the antibiotics feature should be imputed by forward filling, which strategy should be employed and what the sepsis and suspicion windows should be. The necessary features are then extracted from the preprocessed patient data. The preprocessing of [1] includes a normalization in the form of:

$$normalized = \frac{value - mean}{std} \tag{1}$$

Consequently, numerical features are re-normalized and then aggregated per hour. Next, the data is imputed by forward filling. The features for blood cultures, text, mechanical ventilation and catecholamines are excluded by default and the feature for antibiotics is filled depending on what was decided beforehand. Then, the features are cast to their correct type.

E. SOFA

Before the sofa scores can be computed, the Glasgow Coma Scale is computed from its components and the mean arterial pressure is estimated using diastolic blood pressure (DBP) and systolic blood pressure (SBP) according to [16] by:

$$DBP + \frac{SBP - DBP}{3} \tag{2}$$

Now the sofa score is computed for each hour, after which a time of SOFA can be identified according to the guidelines.

F. Suspected Infection and Sepsis Classification

The features for antibiotics and blood cultures are checked according to the set strategy and suspicion window. If the conditions are met, the earlier time is considered to be the time of suspicion, which is then compared to the time of SOFA under the constraints of the sepsis window. As already mentioned, this part of the sepsis check is most critical. If either blood cultures or antibiotics are not reported in the patient data, the patient did not develop sepsis according to the guidelines. During the evaluation of the first experiments, it became clear that, between the suspicion window, strategy and fill procedure, the number one reason for erroneous classifications of patients was the lack of a time of suspicion. At first, the suspicion window was suspected to be the culprit. Increasing the suspicion window to up to ten days increased the suspected infections, however, now, they were suspected too late and missed the sepsis window. This is more serious for [4], because antibiotics need to be administered 72 consecutive hours. If a hospital stay is less than 72 hours, there can be no sepsis. And if the antibiotics data is too sparse, there can be no sepsis without forward filling the feature. Even with forward filling, there can be no sepsis if the patient did not stay at least 72 hours after the first administration of antibiotics.

To tackle this problem, two more strategies were implemented. While the Sepsis-3 strategy uses the first time of blood cultures, and the first time of antibiotics and the supplied suspicion window to compute the time of suspicion, the 'catchsus' strategy takes into consideration all times blood cultures were taken and all times antibiotics were administered. It then checks if any of the possible combinations of time of antibiotics and time of blood cultures fall within the specified suspicion window. For each of the possible combinations that fall within the specified suspicion window, the earlier time is considered the time of suspicion. This can yield multiple times of suspicions, which are then compared to the time of sofa and the sepsis window to generate a sepsis label. The 'grouped' strategy takes this approach even further and expands it onto the time of sofa. As a result, multiple times of suspicion and multiple times of sofa are considered when generating a sepsis label. Unfortunately, even though 'catchsus' and 'grouped' strategies outperformed the standard strategies on unfiltered data – which contains a lot of patients that are impossible to correctly predict for the rule-based guidelines, due to missing antibiotics or blood culture features —, the standard strategies performed better on patients where a positive sepsis label was possible. This indicated that the 'catchsus' and 'grouped' strategies were benefiting from the data distribution rather than being a better strategy.

G. Utilizing Time-Series Forecasting

Next to the obvious benefits of potentially being able to predict the onset of sepsis before the features that would indicate said sepsis are even observed, another advantage of utilizing time-series forecasting is, that the important features can be forecast as well. Unfortunately, in this case, both antibiotics and blood cultures are binary features, whereas [1] is designed to output continuous values. To interpret these continuous values, a threshold was set that assigns everything above or below it a binary label. The method to find said threshold is improvable. So far, a combination of clustering and careful trial and error was used. The experiments that were conducted for this research paper combine observed data and one hour of forecasting output based on that observed data. For each observation window from 20 to 120 hours in steps of four, the resulting concatenation is used as input for the sepsis check.

IV. Data

A. MIMIC-III

Medical Information Mart for Intensive Care 3 (MIMIC-III) is a large database consisting of patients who stayed in the critical care unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. [14] The database consists of twenty six tables. Some examples of the tables include clinical notes, chartevents, admissions and microbiology events. For the full list of tables please refer to Table 4 in [14].

B. Sepsis Label Annotation

Our project is based on the STraTS architecture [1], and as such, we utilized their preprocessing approach to prepare the data in the required format. The authors of [1], however, focus on mortality prediction, which requires a mortality label. In our approach to predict sepsis we need to perform a sepsis check that was introduced earlier in this paper. To identify and label patients with sepsis, ICD9 codes from the diagnosis table have been used. In total, we used 23 codes related to sepsis, as listed in Table VII in the Appendix B. After the data is generated, a sepsis label is assigned to the hospital admission id. Furthermore, we filtered out patients who were admitted with sepsis from our dataset, by parsing the admissions table from MIMIC-III. ICD9 codes are not present in the admission table, therefore, the patients were filtered out based on string matches. These patients have been excluded because the forecasting model cannot benefit by learning from them.

TABLE I: Number of septic/non-septic patients/ICU stays in train/validation/test data.

Data	Non-septic patients	Septic patients	Non-septic ICU stays	Septic ICU stays
Train	26452	2124	33191	3360
Valid	6594	551	8358	904
Test	8296	635	10445	1024

TABLE II: String length and token counts in clinical notes included in our data.

	Avg.	Max.	Min.
String length	1673	55728	3
Num. tokens	316	11336	0

C. Our Data

From MIMIC-III dataset, we built our data from 5288 septic patients (9.2%) and 51994 non-septic patients. We split data into train, validation, test by 64: 16: 20 at patient level (table I).

In addition to the 133 physiological features (see full list in Appendix A), we include 1,407,430 clinical notes from the MIMIC-III dataset in our data. Table II shows text statistics on clinical notes associated with patients included in our data.

D. Clinical Notes Preprocessing

We preprocess clinical notes following common practice for clinical text cleaning by removing stop words and special characters, and normalising case. Additionaly, in order to avoid potential label leakage in the STraTS classification task, we remove sentences containing "sepsis" or "septic".

V. Models

A. STraTS

For both time-series forecasting of physiological feature values for our rule-based sepsis check and direct sepsis label prediction, we use the existing Self-supervised Transformer for Time-Series (STraTS) model [1]. The STraTS model is a deep learning architecture composed of several embedding modules to deal with sporadic clinical time series data. It encodes time-series physiological data in an anti-conventional manner by representing continuous data through a novel Continuous Value Embedding technique to avoid discretizing data (e.g. aggregation, imputation). Each observation in time-series is represented as a triplet: observed time, variable name, and variable value. After initial triplet embedding, the embeddings go through a transformer component with multi-head attention layers to enable learning contextual triplet embeddings, followed by a fusion self-attention module to complete embedding time-series data.

The STraTS architecture supports both a regression model for time-series forecasting and a binary classification model to predict a state label (e.g. mortality, sepsis). The STraTS regression model was originally designed to deal with limited availability of labeled data in the medical domain. It is used during forecasting as an auxiliary proxy task to optimize performance for the classification model. In our case, we fully use both models to obtain forecasting results to support our rule-based sepsis check, and the classification model for 24-hour septic state prediction in a fully data-driven setup.

B. STraTS + Clinical Text Embedding

On the basis of the original STraTS model, which encodes only physiological features as its input for forecasting and classification, we additionally add a clinical text embedding module to the architecture. We embed clinical notes using clinical BERT [5] to obtain text features, and align text features side by side with time-series embedding of physiological features and demographic features for concatenation, and pass through a dense layer to generate outputs for both forecasting and classification task. Figure 1 shows the architecture of the refined STraTS with clinical text embedding.

VI. Results & Discussion

A. STraTS Forecasting

We train STraTS regression models with and without clinical text embeddings to forecast physiological feature values in the two hours following the observation windows, defined as $\{min(0, x-24), x) | 20 \leq x \leq 124, x\%4 = 0\}$. We obtain predictions of both regression models on test data to support rule-based sepsis check. We use masked MSE (mean squared error) for evaluation, where the binary mask indicates if a true value was observed in data.

Table IV shows masked MSE (mean squared error) on test and validation data for STraTS and STraTS + Text regression models. It can be seen from the table that both regression models show similar performance on test data, while the original STraTS without introducing clinical text embeddings had a better MSE on validation data.

B. STraTS Classification

We use 24 hour data (after admission to ICU) to train STraTS classification model with and without clinical text embeddings using random sample of 10, 20, 30, 40, 50 % labelled data to predict septic state for each ICU stay. We repeat each experiment 10 times with different randomly sampled data from train and validation sets.

In the binary classification task, we use three metrics to evaluate model performance on sepsis prediction:

- ROC-AUC: Area under ROC curve.
- PR-AUC: Area under precision-recall curve.
- min(Re, Pr): The maximum of 'minimum of recall and precision' across all thresholds.

Figure 2 shows ROC-AUC, PR-AUC, and min(Re, Pr) of both models evaluated on the test dataset. It can be seen from

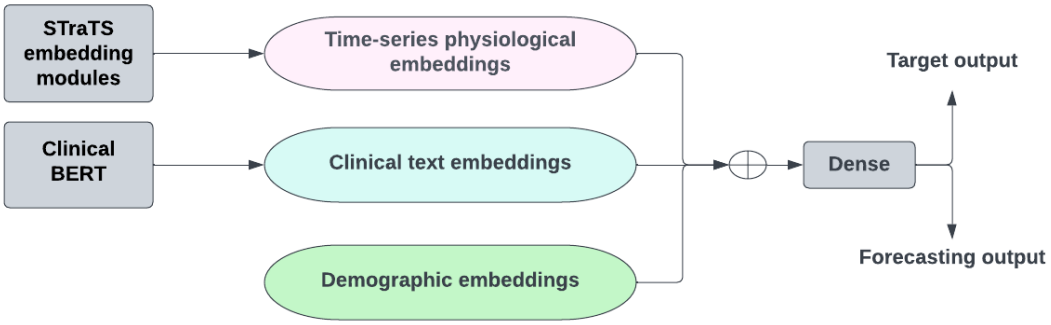

Fig. 1: STraTS + Clinical Text Embedding Architecture.

TABLE III: Sepsis prediction performance on MIMIC-III dataset. The results show mean and standard deviation of the metrics after repeating the experiment 10 times by sampling 50% labeled data each time.

Model	ROC-AUC	PR-AUC	min(Re,Pr)
STraTS	0.891 ± 0.003	0.500 ± 0.009	0.507 ± 0.100
STraTS + Text	0.889 ± 0.002	0.491 ± 0.008	0.492 ± 0.008

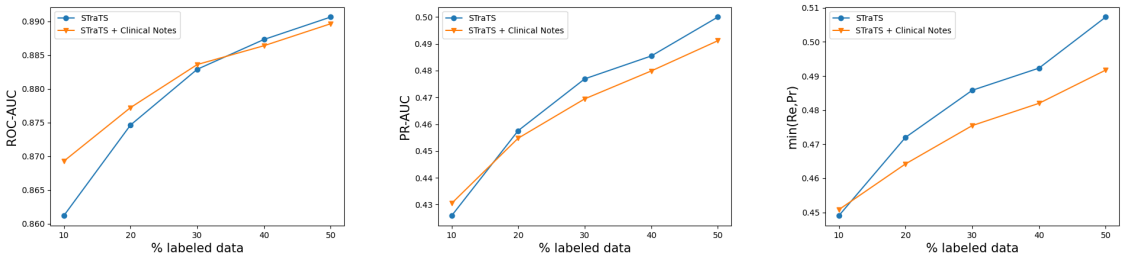

Fig. 2: Sepsis prediction performance on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs.

TABLE IV: Masked MSE (mean squared error) on test and validation data for STraTS and STraTS + Text models.

Model	Test	Validation
STraTS	5.2455	5.2048
STraTS + Text	5.2493	5.5922

the charts that STraTS + Clinical Text only slightly outperforms STraTS when the percentage of labelled data is lower ($\leq 30\%$) in terms of ROC-AUC score. The STraTS model in general has higher PR-AUC and min(Re, Pr) except when only 10% of labelled data is available. STraTS + Clinical Text only shows slightly more advance performance when it in a low-amount labelled data setup, whereas STraTS has better performance with more available labelled data.

C. Sepsis Check results

Since the experiments that utilize time-series forecasting only add one hour of unobserved data, the results in table V are not as significant as we intended. However, it should be noted that the additional data led to more correctly identified true positives in all sets of experiments and even for patients where

TABLE V: 1: experiments were conducted with a suspicion window of 48 and 72 hours, and a sepsis window of 24 and 12 hours. 2: experiments were conducted with a suspicion window of 24 and 96 hours, and a sepsis window of 24 and 12 hours.

Experiment	F1 score	True positives
1.1 observed-only	**0.874505**	251
1.2 observed+forecast	0.873794	**286**
2.1 observed-only	0.87309	234
2.2 observed+forecast	**0.873677**	**265**

the original observed data does not contain both antibiotics and blood culture features. This means that STraTS may enable the sepsis check to overcome the problem of data sparsity. Expanding the STraTS output to multiple hours of predictions seems to be a promising direction for future research.

VII. Conclusion

We implement a rule-based sepsis check and found that the sparseness of patient data is greatly impacting the potential performance. Our rule-based check achieved good performance by itself, and we find in our experiments that by introducing values forecast by the STraTS regression model in its current

form did not improve sepsis check. Additionally, our approach can possibly help to tackle the many complicating real-world factors that can complicate sepsis treatment that are outlined in [7].

On top of the rule-based sepsis check, we use and refine an existing STraTS model [1] for time-series forecasting to support sepsis check and directly predicting sepsis label using 24-hour patient data in a fully data-driven setup. We attempted to improve the STraTS model by integrating a clinical text embedding module based on Clinical BERT to enable multi-modal learning by taking both text and physiological features into account. Both STraTS and STraTS+Text regression models achieved good performance on time-series value forecasting. The classification models also showed high ROC-AUC score in the task of sepsis prediction using features selected in our study. However, the current STraTS+Text model does not outperform the original STraTS model in our experiments. With the physiological time series embedding module working well in both our experiments and in the original work [1], for future work, we intend to improve our model architecture mainly from the text embedding module side through considering possible alternatives such as GloVe [17], BioClinRoBERTa [18], and etc. Additionally, in order to better support the rule-based sepsis check, we plan to further improve the STraTS architecture to enable extended forecasting window based on a short and limited forecasting observation window.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their helpful comments. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

REFERENCES

[1] S. Tipirneni and C. K. Reddy, "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series," 2021. [Online]. Available: https://arxiv.org/abs/2107.14293

[2] K. E. Rudd, S. C. Johnson, K. M. Agesa, K. A. Shackelford, D. Tsoi, D. R. Kievlan *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, no. 10219, pp. 200–211, Jan. 2020. [Online]. Available: https://doi.org/10.1016/s0140-6736(19)32989-7

[3] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 02 2016. [Online]. Available: https://doi.org/10.1001/jama.2016.0287

[4] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati *et al.*, "Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019," *Critical Care Medicine*, vol. 48, pp. 210 – 217, 2019. [Online]. Available: https://physionet.org/content/challenge-2019/1.0.0/

[5] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://www.aclweb.org/anthology/W19-1909

[6] C. M. Torio and R. M. Andrews, "National inpatient hospital costs: The most expensive conditions by payer, 2011," 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK169005/

[7] N. Kissoon, "Sepsis guideline implementation: benefits, pitfalls and possible solutions," *Critical Care*, vol. 18, no. 2, Mar. 2014. [Online]. Available: https://doi.org/10.1186/cc13774

[8] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus *et al.*, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001236921638415X

[9] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. D. Mendonça, H. Bruining *et al.*, "The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, Jul. 1996. [Online]. Available: https://doi.org/10.1007/bf01709751

[10] Y. Wang, Y. Zhao, R. Callcut, and L. Petzold, "Integrating physiological time series and clinical notes with transformer for early prediction of sepsis," 2022.

[11] Z. Liu, L. Wu, and M. Hauskrecht, "Modeling clinical time series using gaussian process sequences," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 623–631.

[12] Z. Liu and M. Hauskrecht, "Clinical time series prediction: Toward a hierarchical dynamical system framework," *Artificial intelligence in medicine*, vol. 65, no. 1, pp. 5–18, 2015.

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[14] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi *et al.*, "Mimic-iii, a freely accessible critical care database," *Nature*, 2016. [Online]. Available: https://www.nature.com/articles/sdata201635

[15] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[16] D. J. Demers and D. Wachs, "Physiology, mean arterial pressure," 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30855814/

[17] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[18] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, Nov. 2020, pp. 146–157. [Online]. Available: https://aclanthology.org/2020.clinicalnlp-1.17

APPENDIX A
FEATURES

TABLE VI: List of 133 physiological features and two static features used in the STraTS-involved workflow.

Physiological Features
ALP
ALT
AST
Albumin
Albumin 25%
Albumin 5%
Amiodarone
Anion Gap
Antibiotics
BUN
Base Excess
Basophils
Bicarbonate
Bilirubin (Direct)
Bilirubin (Indirect)

Bilirubin (Total)
Blood Culture
CRR
Calcium Free
Calcium Gluconate
Calcium Total
Cefazolin
Chest Tube
Chloride
Colloid
Creatinine Blood
Creatinine Urine
D5W
DBP
Dextrose Other
Dopamine
EBL
Emesis
Eoisinophils
Epinephrine
Famotidine
Fentanyl
FiO2
Fiber
Free Water
Fresh Frozen Plasma
Furosemide
GCS_eye
GCS_motor
GCS_verbal
GT Flush
Gastric
Gastric Meds
Glucose (Blood)
Glucose (Serum)
Glucose (Whole Blood)
HR
Half Normal Saline
Hct
Height
Heparin
Hgb
Hydralazine
Hydromorphone
INR
Insulin Humalog
Insulin NPH
Insulin Regular
Insulin largine
Intubated
Jackson-Pratt
KCl
KCl (Bolus)
LDH
Lactate
Lactated Ringers
Levofloxacin
Lorazepam
Lymphocytes
Lymphocytes (Absolute)
MBP
MCH
MCHC
MCV
Magnesium
Magnesium Sulfate (Bolus)
Magnesium Sulphate
Mechanically ventilated
Metoprolol
Midazolam
Milrinone
Monocytes
Morphine Sulfate
Neosynephrine
Neutrophils
Nitroglycerine
Nitroprusside
Norepinephrine
Normal Saline
O2 Saturation
OR/PACU Crystalloid
PCO2
PO intake
PO2
PT
PTT
Packed RBC
Pantoprazole
Phosphate
Piggyback
Piperacillin
Platelet Count
Potassium
Pre-admission Intake
Pre-admission Output
Propofol
RBC
RDW
RR
Residual
SBP
SG Urine
Sodium
Solution
Sterile Water
Stool
TPN
Temperature
Total CO2
Ultrafiltrate
Unknown
Urine

Vancomycin
Vasopressin
WBC
Weight
pH Blood
pH Urine
Demographic Features
Age
Gender

APPENDIX B
SEPSIS CODES FROM MIMIC-III

Table VII contains the sepsis codes from the D_ICD_DIAGNOSES table that were used to determined the positive label for the data.

TABLE VII: ICD9 Codes for sepsis

ICD9 Code	Short Description
0380	Streptococcal septicemia
03810	Staphylcocc septicem NOS
03811	Meth susc Staph aur sept
03812	MRSA septicemia
03819	Staphylcocc septicem NEC
0382	Pneumococcal septicemia
0383	Anaerobic septicemia
03840	Gram-neg septicemia NOS
03841	H. influenae septicemia
03842	E coli septicemia
03843	Pseudomonas septicemia
03844	Serratia septicemia
03849	Gram-neg septicemia NEC
0388	Septicemia NEC
0389	epticemia NOS
67020	Puerperal sepsis-unsp
67022	Puerprl sepsis-del w p/p
67024	Puerperl sepsis-postpart
67030	Puerp septc thromb-unsp
67032	Prp sptc thrmb-del w p/p
67034	Prp septc thrmb-postpart
99591	Sepsis
99592	Severe sepsis

APPENDIX C
STRATS SMALL

We also trained STraTS regression and classification models with data of a smaller set of patients. We used the same 5288 septic patient data and 10555 non-septic patients, resulting in a more balanced dataset (33.4% positive class at patient level). We split data into train/validation/test at patient level (Table VIII) also by 64:16:20. Table VIII shows the number of septic/non-septic patients/ICU stays in the smaller dataset.

Figure 3 shows sepsis prediction performance of STraTS small, STraTS large and STraTS + clinical notes on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs. As shown in 3, STraTS small shows lower ROC-AUC with all percentages of labelled data, whereas it has the highest PR-AUC and min(Re,Pr) compared to STraTS and STraTS + clinical notes, which were trained on the full dataset with more patient data.

APPENDIX D
SEPSIS CHECK COMPONENTS AND VARIABLES

Table X and XI show the components and corresponding variable names within the data for identifying suspected infections and calculating SOFA.

TABLE VIII: Number of septic/non-septic patients/ICU stays in train/validation/test data in the smaller dataset.

Data	Non-septic patients	Septic patients	Non-septic ICU stays	Septic ICU stays
Train	4261	2133	10165	6394
Valid	1071	528	2479	1599
Test	1350	649	3199	1999

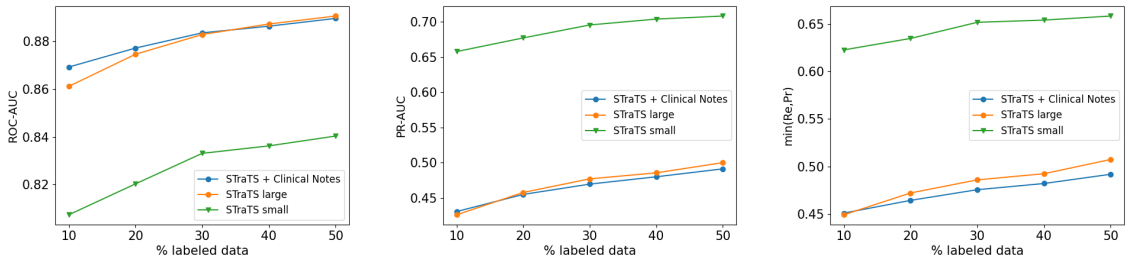

Fig. 3: Sepsis prediction performance on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs.

TABLE IX: Sepsis prediction performance on MIMIC-III dataset. The results show mean and standard deviation of the metrics after repeating the experiment 10 times by sampling 50% labeled data each time.

Model	ROC-AUC	PR-AUC	min(Re,Pr)
STraTS small	0.840 ± 0.003	0.708 ± 0.005	0.658 ± 0.006
STraTS large	0.891 ± 0.003	0.500 ± 0.009	0.507 ± 0.100
STraTS + Text	0.889 ± 0.002	0.491 ± 0.008	0.492 ± 0.008

TABLE X: Suspected infection components and corresponding variable names

component	variable name
time of blood cultures	Blood Cultures
time of antibiotics	Antibiotics

TABLE XI: SOFA components and corresponding variable names

component		variable name
Nervous System	Glasgow Coma Scale	GCS_eye, GCS_verbal, GCS_motor
Cardiovascular	Mean Arterial Pressure	DBP, SBP
	Administration of Vasopressors	Dopamine, Dobutamine, Epinephrine, Norepinephrine
Respiratory System	FiO2 [kPa]	FiO2
	Mechanical Ventilation	Mechanical ventilation
Coagulation	Platelet Count [x10^3/mu l]	Platelet Count
Liver	Bilirubin [mg/dl]	Bilirubin (Total)
Renal	Creatinine [mg/dl]	Creatinine Urine
	Urine [ml/day]	Urine