Original Research

Predicting medical events and ICU requirements using a multimodal multiobjective transformer network

Sudeshna Jana, Tirthankar Dasgupta and Lipika Dey

Research and Innovation—Tata Consultancy Services (TCS), Kolkata 700156, India
Corresponding author: Sudeshna Jana. Email: sudeshna.jana@tcs.com

Impact Statement

Prediction of medical events – such as clinical procedures, susceptibility to adverse reactions, duration of critical care needs, and many others – is essential for providing quality care to patients during their hospital stay. While such predictive models have made use of quantitative variables like patient health data earlier, the present work shows that using qualitative text features from caregiver notes can increase the quality of prediction substantially. The proposed models are multimodal in nature that take into account both quantitative and qualitative variables and generate explainable models. Such models can substantially improve hospital resource management as well as patient care.

Abstract

Effective utilization of premium hospital resources such as intensive care unit (ICU), operating theater (OT), mechanical ventilator, endotracheal tube, and so on plays a significant role in providing high-quality care to critically ill patients within reasonable costs. Non-availability of specialized resources can lead to dire consequences for such patients, and in the worst case, may even turn out to be fatal. However, these resources cannot be kept idle, as they are expensive to maintain. Therefore, one of the core functions of hospital management is targeted at planning and managing these critical resources in order to provide efficient and effective health-care services to the end-users. Predictive technologies play a big role in this. In this article, we present methods for predicting the length of stay in ICU as well as the need for critical interventions for a patient based on the vital signs, laboratory measurements, and the nursing notes of the patient prepared within the first 24 h of ICU stay. The model has been built and cross-validated on the publicly available Medical Information Mart for Intensive Care (MIMIC-III v1.4) data set. We show that the proposed model performs way better than most of the earlier models in the prediction of ICU stay, which had used patient vitals primarily. Experimental results also demonstrate the advantage of using a multiobjective model over independent models for the prediction of ICU stay and critical interventions. The proposed model uses Local Interpretable Model-agnostic Explanations (LIME) that help in identifying the features responsible for predictive decisions. This is very useful in building trust and confidence in the prediction model among clinical practitioners.

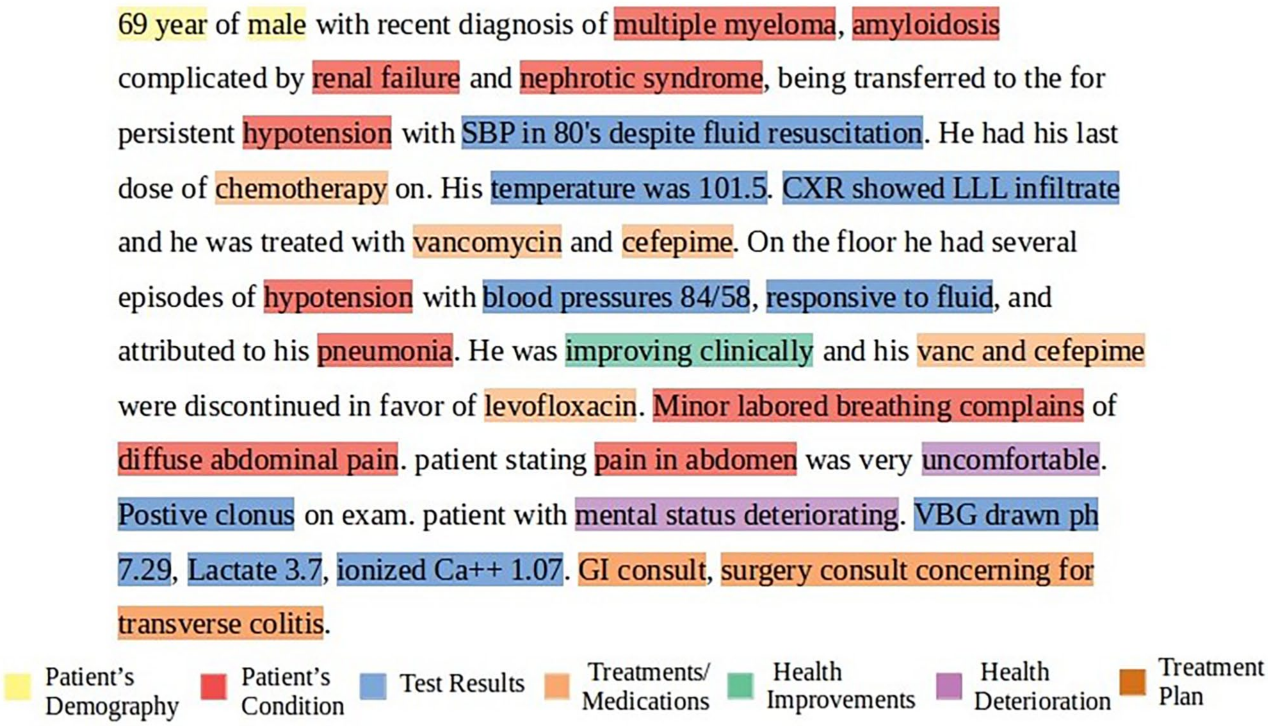
Keywords: ICU length of stay, critical intervention, MIMIC-III, nursing note, severity of illness score, BlueBERT, TF-IDF, LIME

***Experimental Biology and Medicine* 2022; 247: 1988–2002. DOI: 10.1177/15353702221126559**

Introduction

Predicting premium resource requirements for critically ill patients is one of the most important tasks of hospital resource management. Facilities like intensive care units (ICUs) or operation theaters (OTs) are equipped with various expensive monitoring machines, mechanical ventilators, dialysis machines, and so on. Every hospital has only a limited number of these critical resources,[1] and hence, these are known as premium resources. It is also important that such facilities are available to any acutely ill or critically injured patient who may need the highest level of patient care and life support at a given point in time.[2] Patients are referred to ICU if there is a life-threatening deterioration in the patient's condition, or immediately after surgery if the surgery is very invasive and the patient is at high risk of complications. Non-availability of such facilities may lead to fatality.[3] Hence, hospital management needs to do careful planning to ensure that the resources are well-utilized. In this scenario, early prediction of medical events such as clinical procedures, susceptibility to adverse reactions, duration of critical care needs, severity, and so on can help in better management of critical resources, and thereby save patients' lives. Such predictive models can be built using data from Electronic Health Records (EHRs)[4] of past patients. EHR information includes both structured parameters like demographic details, laboratory test results, ward details, and so on and unstructured information like clinical notes which may contain patient history, imaging reports, nursing notes, discharge summaries, and so on. Among these, nursing notes can play a crucial role in data-driven decision-making. Since these time-stamped notes contain the entire history of a patient during hospitalization including assessment of condition, treatment plan, and actions taken. Very few data

ISSN 1535-3702



69 year of male with recent diagnosis of multiple myeloma, amyloidosis complicated by renal failure and nephrotic syndrome, being transferred to the for persistent hypotension with SBP in 80's despite fluid resuscitation. He had his last dose of chemotherapy on. His temperature was 101.5. CXR showed LLL infiltrate and he was treated with vancomycin and cefepime. On the floor he had several episodes of hypotension with blood pressures 84/58, responsive to fluid, and attributed to his pneumonia. He was improving clinically and his vanc and cefepime were discontinued in favor of levofloxacin. Minor labored breathing complains of diffuse abdominal pain. patient stating pain in abdomen was very uncomfortable. Postive clonus on exam. patient with mental status deteriorating. VBG drawn ph 7.29, Lactate 3.7, ionized Ca++ 1.07. GI consult, surgery consult concerning for transverse colitis.

Figure 1. Example of a nursing note for an ICU patient excerpted from our data set. The entities have been highlighted. Conditions are in red, yellow indicates patient's demography, blue indicates test results, treatments/medications are in orange, green indicates health improvement, violet indicates health deterioration, and dark orange indicates treatment planning. (A color version of this figure is available in the online journal.)

sets with detailed EHRs are available for research activities. Most researchers in this area prefer to work with the publicly available Medical Information Mart for Intensive Care (MIMIC)[5] data set.

Early prediction of medical events including predicting the duration of ICU stay has received attention from several research groups in recent times. Barring a few, most of these models were built using only structured information about patients. In 2019, Harutyunyan *et al.*[6] proposed a channel-wise long short-term memory (LSTM) model for predicting in-hospital mortality, physiologic decompensation, phenotype classification, and forecasting the remaining time to be spent in ICU at each hour of stay. The prediction model used 17 vital clinical measurements like capillary refill rate, blood pressure, heart rate, fraction inspired oxygen, Glasgow Coma Scale (GCS), and so on gathered from 4148 ICU stays from the MIMIC database. Subsequently, in 2021, Rocheteau *et al.*[7] revisited the problem and developed a deep learning architecture based on the combination of temporal convolution and pointwise convolution for predicting the remaining ICU length of stay and in-hospital mortality. Their model also utilized only structured time series features like laboratory values, vital signs, and so on from the electronic intensive care unit (eICU) critical care[8] and MIMIC database. Also in 2021, Su *et al.*[9] proposed a model to predict three clinical outcomes such as mortality, severity, and long or short ICU stay for a set of 2224 sepsis patients based on patients' clinical parameters such as age, oxygenation index, white blood cell (WBC) count, oxygen concentration, blood pressure and temperature recorded during the first 6 h in ICU. In a recent work, Alghatani *et al.*[10] reported prediction of ICU length of stay and mortality using several machine learning models from patients' first day's vital signs like heart rate, blood pressure, temperature, respiratory rate, and four demographic features age, gender, height, and weight from the MIMIC data set.

While the clinical parameters are important, the performance of the above-mentioned models suffered as they missed out on valuable information that is contained in EHR texts like patients' medical history, radiology reports, nursing notes, physician notes, and so on. These notes contain more detailed information about a patient's physical and psychological conditions, and treatments prescribed along with observations about a patient's response to treatment. Thus, it can enrich a predictive model by providing additional information about the severity of illness (SOI), signals about observed improvement or deterioration, and details about the treatment plans. For example, linguistic expressions found in these notes like "Lungs are diminished," "patient unresponsive & tremorous," and "unable to determine if ectopy is atrial or ventricular" provide augmented information about the severity of a patient's condition based on expert assessment, that cannot be captured by structured data. Figure 1 shows a sample nursing note with different portions of text color-coded, to highlight the different categories of information that a note may contain.

Neural language models are the default choice for building predictive models that exploit text.[11–15] An end-to-end deep dynamic neural framework was proposed by Pham *et al.*[11] to predict future medical outcomes such as the next diagnosis, current interventions from the current diagnoses, and future risks like unplanned readmission within a certain period. In 2021, Van Aken *et al.*[12] proposed a transformer-based model for predicting multiple clinical outcomes such as the *International Statistical Classification of Diseases and Related Health Problems–Ninth Edition* (ICD-9) diagnosis, ICD-9 procedures, in-hospital mortality, and length of ICU stay, using the discharge summary from the MIMIC data set. Several preprocessing steps were applied to obtain the relevant data. Chrusciel *et al.*[13] proposed the use of random forest along with a word-embedding algorithm based on the Unified Medical Language System (UMLS) terminology, to predict hospital length of stay from unstructured EHRs. Huang *et al.*[14] also investigated the use of physicians and nursing notes generated within the first 48 h of admissions, in predicting ICU length of stay and mortality. They have also utilized the MIMIC database in their study. Very recently, in 2022, Mahbub *et al.*[15] developed a framework for predicting short-, mid-, and long-term mortality based on clinical notes such as ECG, Echocardiogram, and Radiology reports as well as Nursing and Physician notes generated within the first 24 h of admission. Their model also uses data about 37,923 adult ICU patients from the MIMIC database.

A limited number of studies have also explored the use of both structured as well as unstructured data for ICU event predictions. In 2017, Suresh *et al.*[16] proposed a deep neural network model for predicting the onset and weaning of five clinical interventions such as invasive ventilation, non-invasive ventilation, vasopressors, colloid boluses, and crystalloid boluses. These predictions were done every 6 h based on the patient's age, gender, laboratory test values, and clinical notes from the MIMIC database. A convolution-based multimodal architecture was proposed by Bardak and Tan[17] for predicting mortality and length of stay. Medical entities extracted from MIMIC-III clinical notes were used as additional features besides time series ICU signals for the predictions.

Before going into the description of the model proposed in this article, we would like to share some observations about the clinical notes, based on this study. While analyzing the clinical notes for critically ill patients, it is noticed that quite often the first day's nursing notes contain critical inspections about the state of a patient, and do not always hold an accurate estimate of the critical resource requirements for that patient. This study reveals that as critical surgical interventions such as "Bypass," "Stent," and "Tracheotomy" are planned during the course of a treatment, the time involved in preparation and implementation of these interventions, led to a longer ICU stay. This obviously points to the importance of a joint prediction mechanism[18,19] that can effectively predict the duration of ICU stay along with the need for critical interventions as early as possible. Motivated by the above-mentioned facts, in this article, we propose a multimodal, multiobjective framework that does a joint prediction of ICU stay duration and needs for critical interventions. This is a key difference from the earlier models that have mostly considered the problem of predicting multiple outcomes as independent predictive tasks. In the current work, the ICU stay is predicted as "long" or "short" based on whether it is more or less than the average number of ICU length of stay observed for past patients. The predictive model uses the nursing notes and the patient's health parameters recorded during the first 24 h of ICU admission. Prediction of critical interventions can help in determining treatment trajectories early on, effective planning for critical resources, and eventually provide better clinical outcomes.

The proposed model uses both structured data and unstructured text. To represent the nursing notes, we use a transformer-based language model, Bidirectional Encoder Representations from Transformers for clinical texts, BlueBERT[20,21] that is specially trained to model clinical texts. Since transformers can typically encode short texts only, the model also uses an additional Bidirectional Long Short-Term Memory (BiLSTM) network[22] for learning contextualized embeddings of complete nursing notes. We also suggest the use of Term Frequency–Inverse Document Frequency (TF-IDF) feature representations[23] to exploit the presence of medical entities. Along with the text from nursing notes, the model also uses four SOI scores[24–31] that are computed from clinical parameters. These are explained in detail later. In addition, we have utilized a framework called Local Interpretable Model-agnostic Explanations (LIME),[32] which provides human-interpretable insights, in the form of text components like words and phrases, that have contributed most significantly toward the prediction results. This is a key distinguishing aspect of the proposed model, as it makes the model explainable. The model has been evaluated through some detailed experimentation on the MIMIC-III v1.4[5] data set.

The rest of this article is organized as follows. In the next section, we provide a detailed description of the MIMIC-III data set, problem definition, and deep learning models used in this study. Following that, we report the results of our experiments. Finally, we discuss our findings with the interpretability of the predictions and draw a conclusion from our analysis.

Materials and methods

Data source

As our primary data source, we have used MIMIC-III v1.4[5] database, which contains the details of over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. This database has pre-existing Institutional Review Board (IRB) approval, and researchers can access the data by completing the training course "Data or Specimens Only Research" provided by the Collaborative Institutional Training Initiative (CITI).

The MIMIC-III database contains details of 46,520 distinct patients with 58,976 hospital admissions. This database includes both structured and unstructured clinical events documented for patients during hospital admissions. The database is anonymized, and exact dates and times of events have been obfuscated.

Data extraction and preprocessing

In this study, we have included patients of age 18 years and older only, who were admitted to the ICU at BIDMC from 2001 to 2012. Patients in the younger age group usually did not need unplanned critical interventions. During a single

Table 1. List of all the clinical records used in the experiment.

Unstructured	Nursing notes
Structured	Age, pre-ICU length of stay, mean arterial blood pressure, systolic blood pressure, temperature, heart rate, respiratory rate, sodium, potassium, hematocrit, WBC, creatinine, GCS platelet, PaO_2, oxygenation, pH, urea, bicarbonate, bilirubin, urine output

ICU: intensive care unit; WBC: white blood cell; GCS: Glasgow Coma Scale.

hospital admission, a patient could have undergone multiple ICU admissions. In this work, we consider only those admissions which have a single ICU stay with a duration of more than 24 h. We have excluded the admissions which have no nursing notes available in the first 24 h. After applying all the criteria discussed above, we are left with 28,659 unique hospital admissions in our data set.

For each admission, we have extracted all the nursing notes that were recorded within the first 24 h of ICU admission and concatenated them to obtain a single note. All notes have been converted to lowercase, and non-alphanumeric characters have been discarded. In addition to the textual features, we have selected 20 vital signs and lab measurements available in the first 24 h of ICU stay, which are presented in Table 1. For features that have multiple values recorded within the first 24 h, only the worst value is considered. Like all such data sets, these data also suffer from missing values. For the proposed model, the missing values are filled up by carrying forward the last observation, when available; if no previous observation is available, the value is imputed with a physiologically normal value. Instead of directly utilizing these structured features, we have used four different scores that are computed from the clinical parameters and are indicative of the SOI. These are as follows: (1) Acute Physiology and Chronic Health Evaluation (APACHE-II) score,[24–26] (2) Simplified Acute Physiology Score (SAPS-II),[26,27] (3) Sepsis-related Organ Failure Assessment (SOFA) score,[28,29] and (4) Oxford Acute Severity of Illness Score (OASIS).[30,31] These scores reflect the degree of illness, complexity of the disease, and degree of organ system derangement for the ICU patients.[33–35]

The APACHE-II score is calculated based on age, chronic health status, and 12 physiological variables that include mean arterial pressure, temperature, heart rate, respiratory rate, oxygenation, GCS, pH, sodium, potassium, creatinine, hematocrit, and WBC level in the blood. SAPS score uses logistic regression techniques to predict the SOI using 12 physiological variables, age, type of admission such as surgical or medical, and three variables related to acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy. The SOFA score is used to measure a person's organ function or rate of failure during the stay in an ICU. This score is based on six different values coming from the assessment of the respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. The OASIS score is computed from 10 variables: elective surgery, age, pre-ICU length of stay, and seven physiological measurements. For our model, all four scores are calculated using data collected in the first 24 h of ICU stay only. Finally, the processed nursing note and the SOI scores serve as inputs for our prediction model.

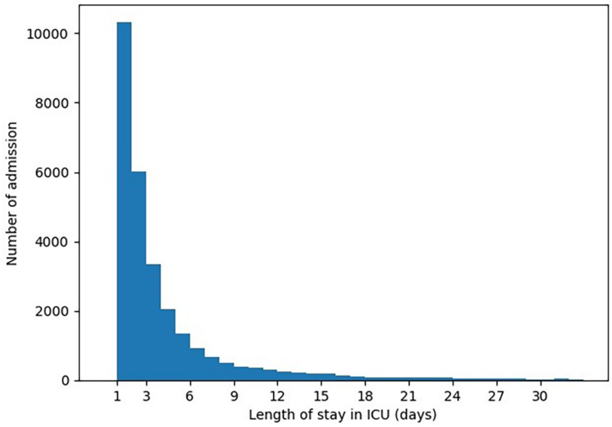


Figure 2. Histogram showing the ICU length-of-stay distribution of 28,659 hospital admissions. (A color version of this figure is available in the online journal.)

Data labeling

As we have discussed above, in this work, we mainly focus on two vital clinical prediction tasks, length of stay, and requirements of critical interventions in the ICU. The explanation of these tasks and the class distributions are as follows:

Length-of-stay prediction. Rather than predicting the exact number of days in ICU, which can vary due to operational reasons, our model predicts whether the length would be short or long, depending on whether it is less or more than a threshold. The threshold is decided as the median of past stays observed in a hospital. This way the model can be easily adapted for different types of hospitals and regions very easily. As shown in Figure 2, the median ICU length of stays is 3 in our data set, so we have defined a threshold of three days. All admissions in the data set were labeled with a categorical value which is "Short" if the ICU length of stay <3 days and "Long" if the ICU length of stay $\geqslant 3$ days.

Intervention prediction. Interventions are either a procedure or treatment applied to a patient to prevent the patient's condition from deteriorating. Although the proposed architecture is generic enough to predict any critical intervention requirement for critically ill patients along with the duration of the ICU stay. In this study, we have worked on predicting whether any of the following four surgical interventions—*coronary artery bypass surgery*, *stenting*, *tracheotomy*, and *cholecystectomy* will be required for a patient or not. These interventions were picked up as they were the most commonly occurring surgical interventions in our data set.

Coronary artery bypass surgery is performed to treat a blockage or narrowing of one or more of the coronary arteries and restore the blood supply to the heart muscle.[36] Stents are also often used to treat narrowed coronary arteries that provide the heart with oxygen-rich blood. Doctors usually recommend bypass surgery over stent implantation for patients who have severe coronary artery disease and multiple blockages.[37] Stenting is minimally invasive, so the recovery is usually easier than it is with bypass. Other than coronary artery disease, stents are also used to treat blocked airways, aortic aneurysms, biliary paths, and so on. It was also observed that tracheotomy is often needed for ICU patients, whenever the usual route for breathing is somehow blocked for a patient.[38] Tracheotomy provides an air passage by creating a hole at the front of the neck so a tube can be inserted into the trachea which helps patients in breathing. A cholecystectomy is a common surgical intervention to treat gallstones and the complications they cause.[39]

Our first task was to generate correctly labeled data to build the predictive model. Every record in the data set that is to be used for training or testing needs five labels. The first one is categorical and takes value short or long to denote the length of ICU stay. The remaining four are binary in nature – 1 indicating the presence of an intervention, and 0 denoting its absence. In the MIMIC-III database, there is no specially assigned label for these interventions. To overcome this barrier, we have designed a language processing task to assign these labels to the admission records. It was observed that the discharge summaries contained very detailed information about a patient, the illness, and the treatment under different heads like "past medical history," "Present history of illness," "brief hospital course," 'treatments provided," "list of all medications," and so on. It is also observed that if an intervention has been implemented for a patient, it is almost invariably mentioned in the discharge note. We used language tools called Named Entity Recognizers (NERs) that can detect different types of entities from text documents. Our intent was to identify the names of interventions from the discharge summary.

The detection and classification of intervention entities from the discharge summary were done using Named Entity Recognition (NER) architecture as proposed by Devlin *et al.*[40] while introducing transformer-based language models for language processing tasks. We have fine-tuned this model with the EBM-NLP corpus,[41] which contains 5000 annotated abstracts of medical articles with names of Patient population enrolled, Interventions, and the Outcomes measured (PICO) marked. The articles contain details of patients' demography, health conditions, lifestyle along with mentions of seven types of interventions such as surgical, physical, pharmacological, psychological, educational, control, and others. We have also replaced the underlying BERT model with BlueBERT-Base-uncased.[20] For fine-tuning, we have used the Adam optimizer with a learning rate of $5e^{-5}$ and a batch size of 32. The task of the model is to learn a label for each word in a sentence. The label denotes whether a word is part of an intervention name or not and if it is then the right intervention category. Figure 3 shows the proposed architecture along with the labels assigned to the words from a sample sentence.

Later, this fine-tuned model is run on discharge summaries to extract the names of all interventions contained in them. For each admission, if any of the above-mentioned surgical interventions are present in the discharge summary, we have considered the label of this surgical intervention as "1" otherwise "0."

Table 2 represents the detailed description of the data set thus created, and lists the sample size of each class. In Figure 4, we have shared the data for the clinical interventions mentioned in the first day's nursing notes versus their mentions in the later day's notes. It can be seen that for around half of the patients who underwent bypass surgery or stenting, these interventions were not mentioned in the first day's nursing notes, indicating that these were not planned. For "Tracheotomy" and "Cholecystectomy," this percentage is even lower, which shows that most of these interventions were unplanned. We have also observed in our data set, that in several cases, multiple interventions were performed for a single patient. Figure 5 presents detailed data about all such cases of multiple interventions in the form of co-occurrence networks. This further justifies the advantage of designing the proposed predictive model as a multiobjective one, since clearly the interventions are not mutually exclusive. For our experiment, we have split the data set randomly into 80-10-10, with 80% data serving as the training set (22,928 admissions), 10% as the validation set (2866 admissions), and 10% as the test set (2866 admissions).

Model development

In this section, we present the details of the architecture of our proposed multimodal multiobjective method. Figure 6 presents the architecture. The model consumes three different types of inputs. The nursing notes are converted into contextualized vector representations using BlueBERT. A TF-IDF vector is computed using the important features from nursing notes. The SOI scores are generated using publicly available tools.

Input preparation

Transformer-based representation of nursing notes. BERT[40] is a pretrained language representation based on the transformer encoder architecture, that has exhibited outstanding performance in various NLP tasks like question-answering, summarization, inferencing, and so on. The representation being contextual, the language model outperforms earlier methods that used Word2Vec embeddings or GloVe representations.[42] It is also known that fine-tuning BERT on appropriate literature further improves the performance of downstream tasks. BlueBERT[20] is one such model that is pretrained over a corpus of biomedical research articles sourced from PubMed abstracts and MIMIC-III clinical notes. Therefore, we have used BlueBERT for creating the contextual representation of the nursing notes. The pretrained model has 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads. The input of the BlueBERT model is represented as token embedding, a learned segment embedding for identifying the sequence of the token, and position embedding corresponding to the token's position in the input sequence. A classification token

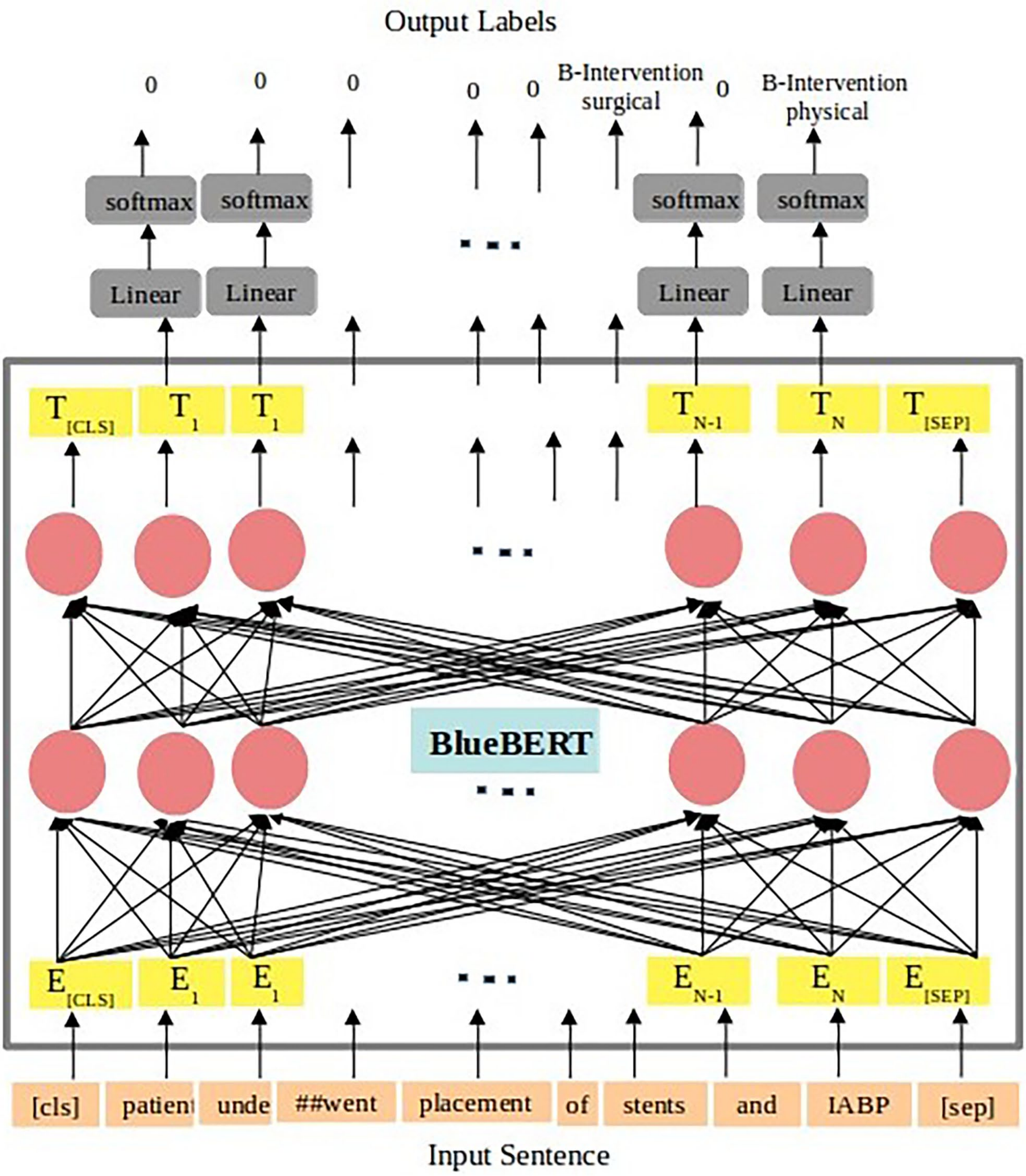


Figure 3. The architecture of the proposed BlueBERT model for Named Entity Recognition. (A color version of this figure is available in the online journal.)

(CLS) is inserted in the front for classification tasks and an (SEP) token at the end of the sequence of input tokens.

The limitation of the BlueBERT model is that it cannot handle texts which are longer than 512 tokens. In our data set, the average number of tokens per nursing note is 2000. To obtain a representation for the entire note, it was split into multiple chunks of 400 tokens each, with 50 overlapped tokens between two consecutive chunks. Each chunk is fed to the BlueBERT model, and the output from the last transformer layer is retained to be used by the next layer to create a note-level representation. For capturing the long-term relationships among the chunks of a single note, we have added a BiLSTM layer clubbed with an attention layer over the BlueBERT representation,[22] as shown in Figure 6. The output vector of each chunk from the BlueBERT model is fed as input to the BiLSTM layer with 100 units. The attention layer on top of it helps to learn class-word correlations. Using this architecture, it is possible to use notes of arbitrary

Table 2. Summary of the data set used for ICU stay classification and intervention prediction task.

Total number of admissions	28,659
Maximum number of tokens in a note	5331
Average number of tokens in a note	2000
Classes for ICU length-of-stay prediction	2 (“Short,” ICU length <3 days and “Long,” ICU length ⩾3 days)
Classes for intervention prediction	4 (“Bypass,” “Stent,” “Tracheotomy,” “Cholecystectomy”)
Sample size of the class “Short”	16,321
Sample size of the class “Long”	12,338
Sample size of the class “Bypass”	6343
Sample size of the class “Stent”	4078
Sample size of the class “Tracheotomy”	1537
Sample size of the class “Cholecystectomy”	1065

ICU: intensive care unit.

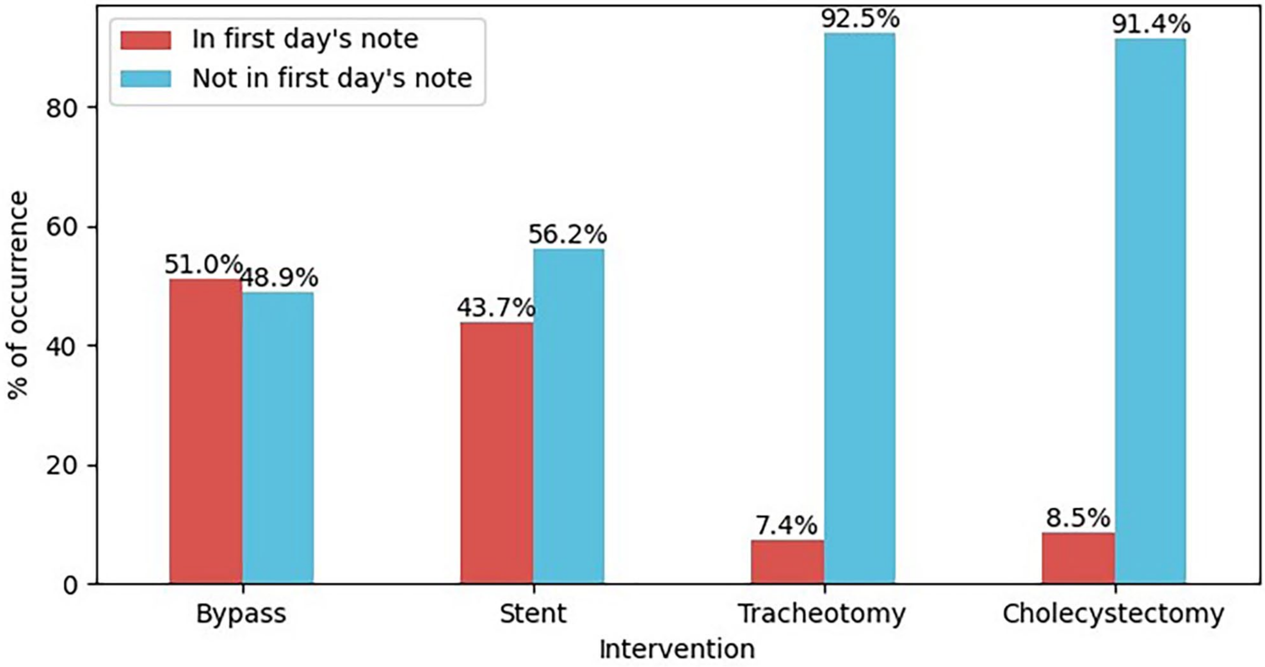


Figure 4. Percentage of occurrences of four interventions in the first day’s nursing note versus notes recorded on other days. (A color version of this figure is available in the online journal.)

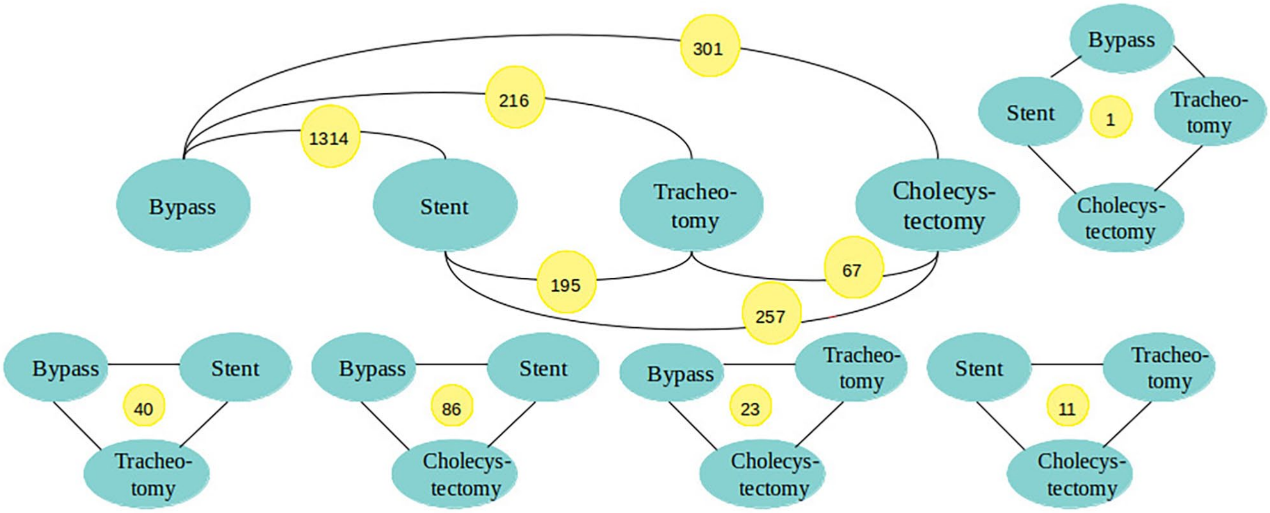


Figure 5. Co-occurrence networks of four interventions in our data set. The nodes of these graphs are the interventions and the values in the yellow circle represent the number of admissions in which all these interventions happened together for the patients during their ICU stay. (A color version of this figure is available in the online journal.)

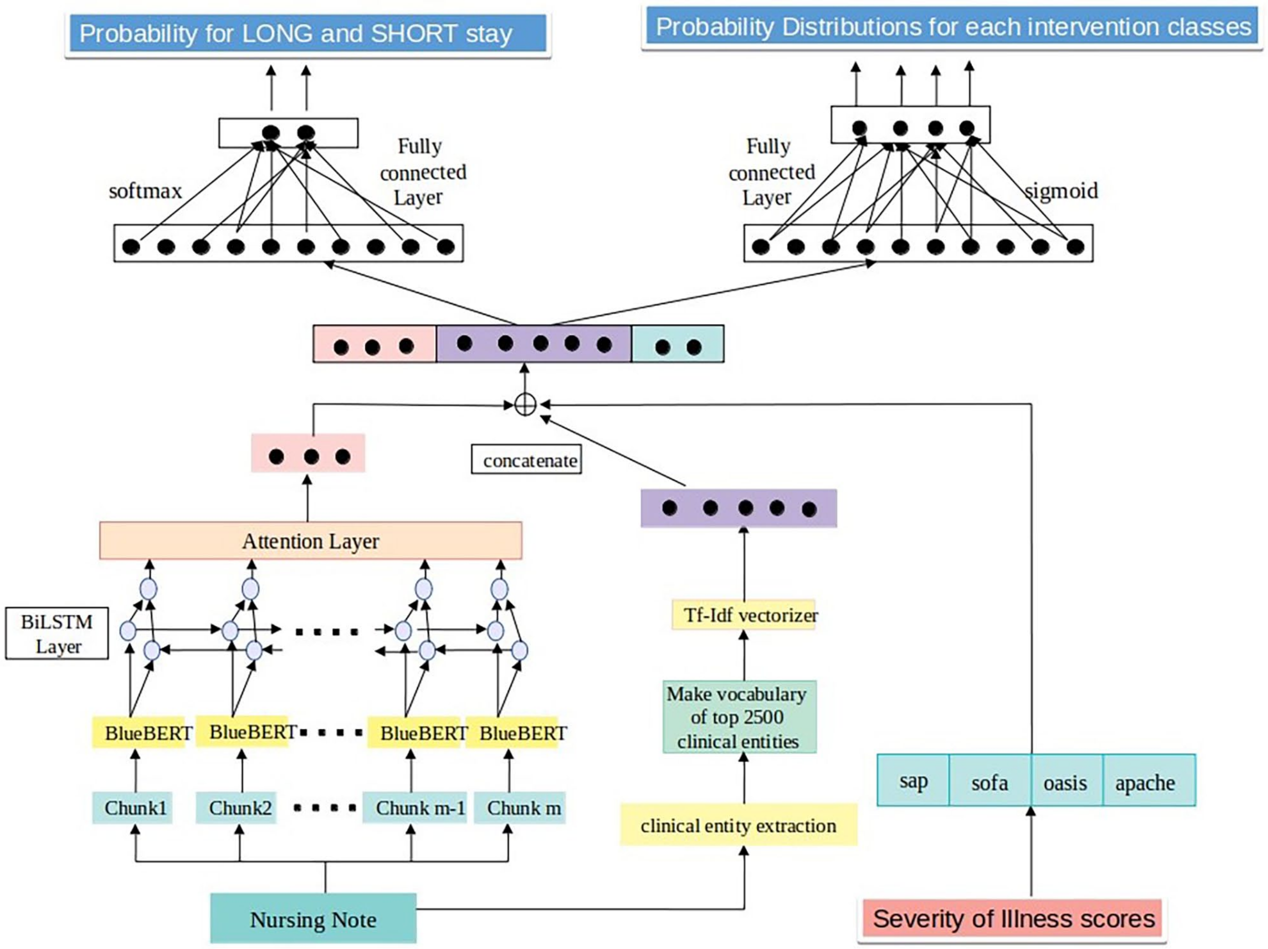


Figure 6. Overview of proposed multimodal multitask framework for predicting the ICU length of stays and necessity of the interventions. Process the nursing notes in chunks by the BlueBERT model and add a BiLSTM-attention layer on the top. We also extract 2500 medical entities from these notes and make a TF-IDF representation. Then, the note representations from the BlueBERT–BiLSTM-attention network, TF-IDF representation, and four severity of illness scores are concatenated, and two task-specific fully connected layers are applied to obtain the final predictions. (A color version of this figure is available in the online journal.)

length as inputs. Since the number of chunks needed to represent a single note is not fixed, we process the note in batches. Each batch accommodates 5 chunks of 400 tokens each. A note may come in a single batch or multiple batches, with padding for shorter sequences, whenever necessary. Finally, the output of the BiLSTM-attention network serves as an embedded representation of the nursing note.

Entity-class relationship captured using TF-IDF measure. Clinical Named Entities like names of drugs, diseases, treatments, and so on present in a text can be quite indicative of its class. To make use of this additional aspect, we have utilized the frequency-based feature weighting mechanism that was quite popular for document classification tasks in the preneural era. One such measure is the TF-IDF[23] value, that captures the relative frequency of a term in a single document versus its global presence in the repository. The TF-IDF score for each term is defined as follows

$$\text{TF-IDF}\left(t,d\right)=\text{TF}\left(t,d\right)\times\text{IDF}\left(t\right)$$

where TF(t, d) is the number of times term "t" occurs in a note d, and $\text{IDF(t)}=log(n/(df(t)+1))$, where n is the total number of notes and $df(t)$ is the number of notes in which term t occurred.

We have added TF-IDF value for clinical named entities extracted from nursing notes like "chest pain," "abdominal pain," "respiratory distress," "blood loss hemorrhage," and many others, as an input feature for the prediction model. A comprehensive list of 2500 most frequent clinical entities present in the data set was obtained using two different NER tools. One is CliNER,[43] which is an open-source natural language processing system for extracting clinical concepts such as diseases/disorders, treatments/medications, and tests from a clinical text. The other one is the BlueBERT-based NER model fine-tuned with the EBM-NLP corpus that we have implemented, as described earlier. A 2500 dimensional feature vector was created for each note using the TF-IDF score of each entity computed for that note.

SOI scoring. As described earlier, in our work, for each admission we computed four types of SOI scores: APACHE-II, SAPS-II, SOFA, and OASIS based on data collected within 24 h of ICU admission. The ranges of these scores are all different. APACHE-II score lies between 0 and 71, SAPS-II score ranges from 0 to 163 points, the range of SOFA score

Table 3. Results of our experiments for ICU length-of-stay classification task using different performance metrics.

Model	Accuracy	F1 score 'Short'	F1 score 'Long'	AUC
Multitask BiLSTM-blueBERT with tf_idf and SOI	**0.84**	**0.86**	**0.82**	**0.89**
Multitask BiLSTM-blueBERT with tf_idf	0.83	0.84	0.80	0.87
Multitask BiLSTM-blueBERT	0.79	0.79	0.78	0.85
BiLSTM-blueBERT with tf_idf and SOI	0.79	0.81	0.79	0.86
BiLSTM-blueBERT with tf_idf	0.79	0.80	0.78	0.84
BiLSTM-blueBERT	0.77	0.77	0.76	0.83

ICU: intensive care unit; AUC: area under the receiver operating characteristic; SOI: severity of illness.
We report accuracies, AUC scores of the model, and F1 scores of both the classes "Short" and "Long." Bold values indicate the best performance of our experiments.

is 0–24, and then OASIS score ranges from 6 to 64. We have normalized all the scores by Min–Max normalization technique and ensured that all the scores lie between 0 and 1.

The final input representation is a 1 × 2704 dimensional vector generated by concatenating the outputs of the BlueBERT–BiLSTM-attention layer, the TF-IDF feature vector for clinical entities, and the SOI scores.

Prediction network. The concatenated vector embedding is fed into two task-specific fully connected layers, one for predicting ICU length-of-stay and another for predicting the possibilities of each of the four surgical interventions—bypass surgery, stenting, tracheotomy, and cholecystectomy. For the length-of-stay classification, we have used softmax activation function with binary cross-entropy loss L_{LOS}. For the intervention prediction tasks, since these are not mutually exclusive outcomes, we have trained the prediction layer using sigmoid activation function with binary cross-entropy loss functions $L_{Intervention}$. These two loss functions are defined as follows

$$L_{LOS} = \frac{-1}{N}\sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i)$$

where N defines the batch size, y_i is the actual label, and p_i is the softmax probability for ith data in length-of-stay classification task

$$L_{Intervention} = \sum\nolimits_{s \in interventions} L_s$$

where L_s is the binary cross-entropy loss for prediction of surgical intervention s, and it is defined as

$$L_s = \frac{-1}{N}\sum_{i=1}^{N} y_i^s log(p_i^s) + (1 - y_i^s) log(1 - p_i^s)$$

where N defines the batch size, y_i^s is the actual label, and p_i^s is the probability for ith data for the surgical intervention s.

Finally, we have defined a joint loss function using a linear combination of the loss functions for the two tasks as

$$L_{joint} = \lambda * L_{LOS} + (1 - \lambda) * L_{Intervention}$$

where λ controls the contribution of losses of the individual tasks in the overall joint loss. We have minimized this joint loss function with the Adam optimizer with an initial learning rate of 0.001 for training. Our experiment was implemented in Pytorch. The batch size and the sequence length choices are guided by the available GPU memory.

Results

We now present results from our experiments to predict ICU length of stay and the possibilities of four interventions based on observations made during the first 24 h of ICU admission for a patient. We have used the ablation mechanism, whereby the performance of the proposed model is compared with simpler models that either use fewer types of inputs or are not multiobjective in nature.

To measure the performance, we have used three different evaluation metrics: accuracy, F1 score, and area under the receiver operating characteristics (AUROCs or AUCs). The accuracy score is computed as the percentage of correct predictions made for a test data set. Although accuracy is a standard measure for reporting classification performance, high accuracy scores can be misleading without information about how true positive and true negative distribute across the data set, which is an important evaluation for medical data analytics. Since the intervention prediction task suffers from a class imbalance problem, the accuracy score is not enough for evaluating model performance. F1 score is the metric that calculates the harmonic mean of precision and recall, where precision is the ratio of correctly predicted positive observations to the total predicted positive observations and recall is the ratio of correctly predicted positive observations to the actual positive observations. Therefore, this score takes both false positives and false negatives into account and provides a better assessment of model performance. The AUC is a popular robust metric for highly imbalanced data sets, and it computes the AUROC curve. The ROC curve shows the trade-off between true-positive rate (TPR) and false-positive rate (FPR) and provides the ability of a classifier in distinguishing between classes. In particular, we have considered macro AUC for our intervention prediction task. The macro AUC is defined as the average of the per class AUCs.

Tables 3 and 4 summarize the overall performance of our experiments for the two prediction tasks. The first rows in each table show the performance of the proposed model, while the other rows present the performance of simpler models. Single objective BiLSTM–BlueBERT-attention with TF-IDF and SOI model developed for the two tasks independently could attain an accuracy of 0.79 for ICU length-of-stay prediction and 0.80 for intervention predictions. We observe that the use of dependencies between two tasks significantly

Table 4. Results of our experiments for intervention prediction using different performance metrics.

Model	Accuracy	F1 score "Bypass"	F1 score "Stent"	F1 score "Tracheotomy"	F1 score "Cholecystectomy"	AUC
Multitask BiLSTM-blueBERT with tf_idf and SOI	**0.82**	**0.89**	**0.83**	**0.55**	**0.54**	**0.86**
Multitask BiLSTM-blueBERT with tf_idf	0.81	0.86	0.81	0.53	0.51	0.85
Multitask BiLSTM-blueBERT	0.80	0.85	0.78	0.51	0.48	0.83
BiLSTM-blueBERT with tf_idf and SOI	0.80	0.85	0.79	0.49	0.49	0.83
BiLSTM-blueBERT with tf_idf	0.78	0.81	0.78	0.48	0.46	0.83
BiLSTM-blueBERT	0.77	0.81	0.79	0.47	0.46	0.81

AUC: area under the receiver operating characteristic; SOI: severity of illness.
We report accuracies, AUC scores of the model, and F1 scores of all four classes. Bold values indicate the best performance of our experiments.

Table 5. Behavior of our best model on the test data set in four interventions prediction tasks.

	Bypass	Stent	Tracheotomy	Cholecystectomy
Total number of occurrences	615	401	157	97
Doctors recommend in first day's note	308 (50.1%)	174 (43.4%)	5 (3.1%)	10 (10.3%)
Our model correctly recommend intervention from first day	603 (98%)	302 (75.3%)	112 (71.3%)	40 (41.2%)
Number of false negative cases	12 (2%)	99 (24.6%)	45 (28.6%)	57 (58.7%)

improves the performance over other models. The length-of-stay classification accuracy improved from 0.79 to 0.84 (0.86–0.89 AUC), while the intervention prediction accuracies also improved from 0.80 to 0.82 (0.83–0.86 AUC). The advantage of using all the features is also obvious from the tables. Using the TF-IDF information along with the clinical notes improved the accuracy of the predictions from 0.79 to 0.83 for ICU length-of-stay prediction, and from 0.80 to 0.81 for intervention prediction tasks. Adding the SOI scores along with them further improves the prediction accuracies to 0.84 for ICU length-of-stay prediction and 0.82 for intervention prediction. In addition, we have observed a significantly high F1 scores were obtained for bypass surgery (0.89) and stenting (0.83), which were more prolific in the data set. Lower F1 scores for tracheotomy and cholecystectomy may be attributed to the smaller sample sizes as shown in Table 2.

Discussion

Our results indicate that all models performed better when we utilized unstructured nursing notes in combination with structured clinical measurements. The best performance was obtained by jointly modeling the two prediction tasks. This is not very unexpected as analysis reveals that for 52% of bypass surgery patients and 94% of tracheotomy patients, the predicted ICU length of stay is long. Table 5 contains a detailed analysis of the model performance for the prediction of the four interventions. It is important to note that for only 50% of the patients who underwent a bypass surgery, doctors had suggested it on the first day of ICU admission as reflected in the first day's nursing note. The proposed model predicts this intervention for many more patients based on the first day's nursing notes, from among which 98% turned out to be true positives. This establishes the power of predictive models in treatment planning. Clearly, the predictive model that is built from large volumes of past data can link similarities across notes and predict the interventions even before the medical experts.

The TPRs for predicting "Stenting" and "Tracheotomy" are 75.3% and 71.3%, respectively. Interestingly, out of a total of 157 patients who underwent tracheotomy in the test set, doctors had predicted it on the first day for five patients only, which is 3% of the total number. The proposed model could correctly predict the need for "Tracheotomy" based on the first-day nursing notes for 112 patients (71.3%). The predictive power of the model comes from the ability to connect phrases like "respiratory failure" or "pneumonia" to the intervention class tracheotomy, even before the attending doctors prescribe it.

Interestingly, although at first glance the performance of cholecystectomy prediction may not seem to be good, careful analysis shows that out of 97 cholecystectomy cases that exist in the test data set, doctors had suggested this intervention on the first day for 10 patients only. The proposed prediction model has correctly predicted this intervention for 40 cases from their first day's nursing notes. Most of these nursing notes contained phrases like "abdominal pain," "Gastrointestinal bleeding," and "pancreatitis."

We have also done an analysis of the false positives or the wrong predictions for different interventions. In 4.7% of cases, our model wrongly predicts "Bypass." Analysis reveals that most of these nursing notes contained phrases like "Aortic Stenosis," "Mitral Stenosis," "Severe Chest Pain," and "Heart Block." This is not totally absurd. Doctors still consider it medically reasonable, because usually bypass is required for these patients. For 2% of the false-negative cases that the model missed recommending "Bypass" surgery, it was found that these people were admitted with diseases like "Fever," "Hypotension," "Dyspnea," and "Unstable Angina," and there was no sign of severity in day 1. Similarly, it was found that most of the false-positive predictions for stenting were for patients who were suffering from "chest pain" and "breathing difficulties." Also, sometimes the stent was needed during an emergency procedure to open the blocked coronary artery, clearly for those cases, model could not predict a stent on the first day. In 4.3% of admissions, our

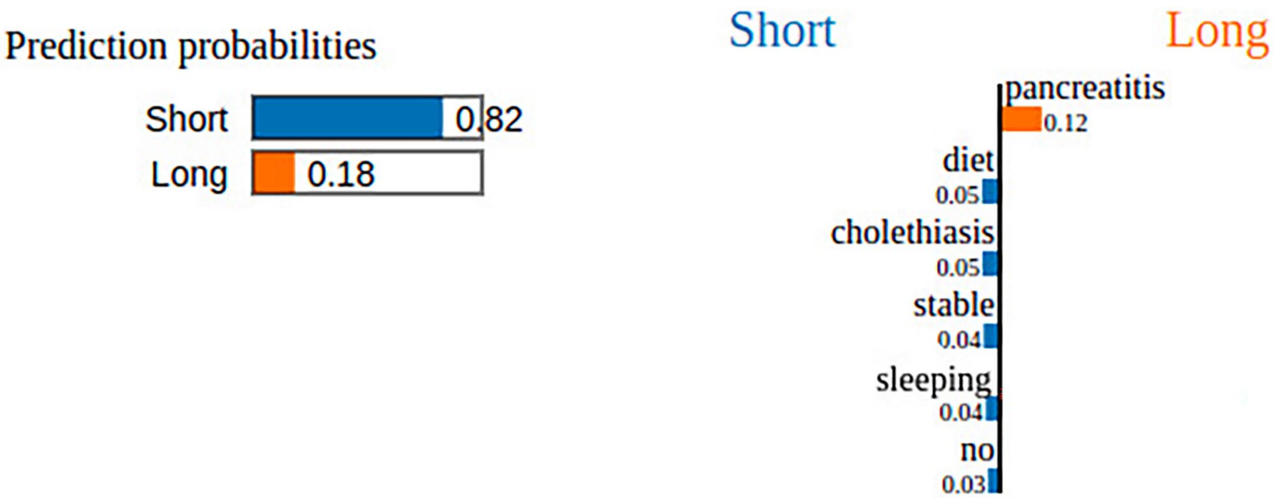


Figure 7. Example of an LIME explanation for prediction probabilities of our model for ICU length-of-stay prediction task where actual: “Short” and predicted: “Short.” The bars’ length highlights the specific contribution of each word of the nursing note: the blue ones push the model toward “Short” prediction, whereas the orange ones to “Long.” (A color version of this figure is available in the online journal.)

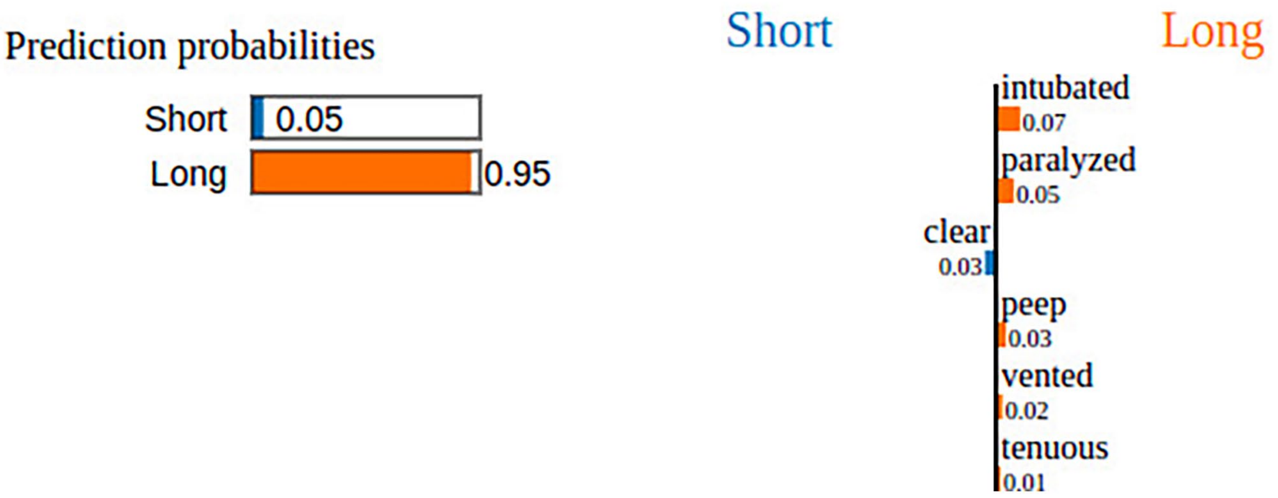


Figure 8. Example of an LIME explanation for prediction probabilities of our model for ICU length-of-stay prediction task where actual: “Long” and predicted: “Long.” The bars’ length highlights the specific contribution of each word of the nursing note: the blue ones push the model toward “Short” prediction, whereas the orange ones to “Long.” (A color version of this figure is available in the online journal.)

prediction model wrongly predicted “Tracheotomy,” and for very few admissions such as 0.3%, model wrongly predicted “Cholecystectomy.”

The phrase-based explanations that we have presented here were obtained using an explanation generation framework called LIME.[35] Explainability plays a crucial role in the health-care domain as it helps the model gain trust of medical practitioners and other health-care professionals. While the classification performance of a model is important, it is also crucial for them to understand its underlying decision-making process. Medical practitioners like to know the parameter values that led to a predictable outcome. Such insights also help in identifying any inherent incorrect bias that might inadvertently creep into the model due to the data used.

LIME is a local surrogate interpretable framework, which can be applied to any black box model to generate explanations and insights. It conducts tests on single data elements by tweaking the feature values and observes the resulting impact on the output. We have applied “LimeTextExplainer” on randomly selected text from each class to generate local explanations for predictions and understand the prediction strategy of our model. For creating LIME output, we define the explanation as “explainer.explain_instance,” which shows the calculated prediction probability of classes and the six most influential features with their weights that have influenced the predictions. As we can see from the example in Figure 7, the LIME explainer emphasized the word “diet,” “stable,” “sleeping,” and so on, while predicting the class of “Short” stay for the note shown in the image. The LIME output of a note for “long” stay is presented in Figure 8, and the highest weighted features are found to be “intubated,” “paralyzed,” “vented,” and so on, which have been learnt as major indicators of long-stay during training. Figures 9 to 12 depict the LIME outputs for nursing notes corresponding to the intervention classes “Bypass,” “Stenting,”

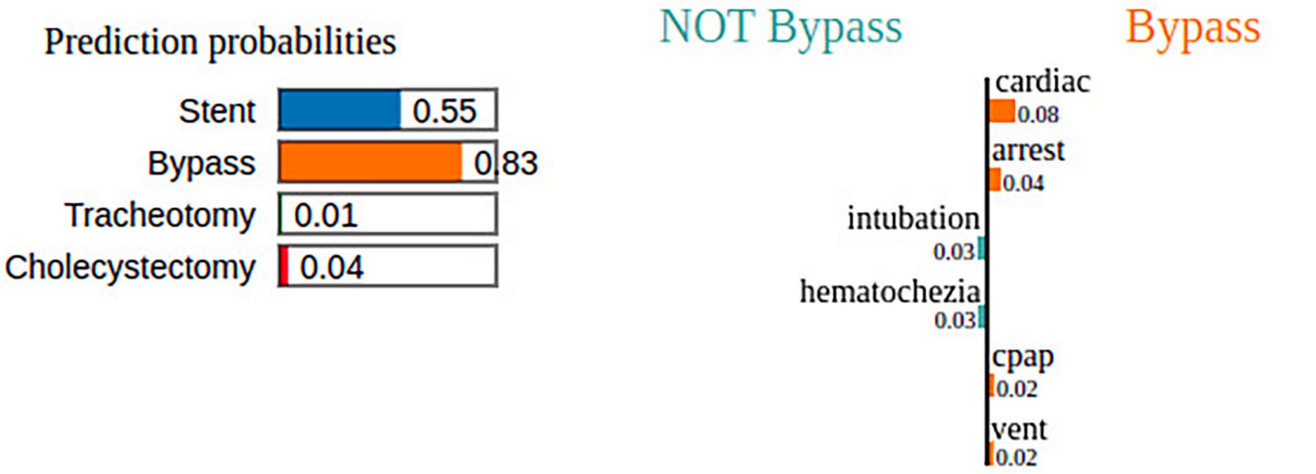


Figure 9. Example of an LIME explanation for prediction probabilities of our model for intervention prediction task where actual: "Bypass" and predicted: "Bypass." The bars' length highlights the specific contribution of each word of the nursing note: the orange ones push the model toward "Bypass" prediction, whereas the other colors for other interventions. (A color version of this figure is available in the online journal.)

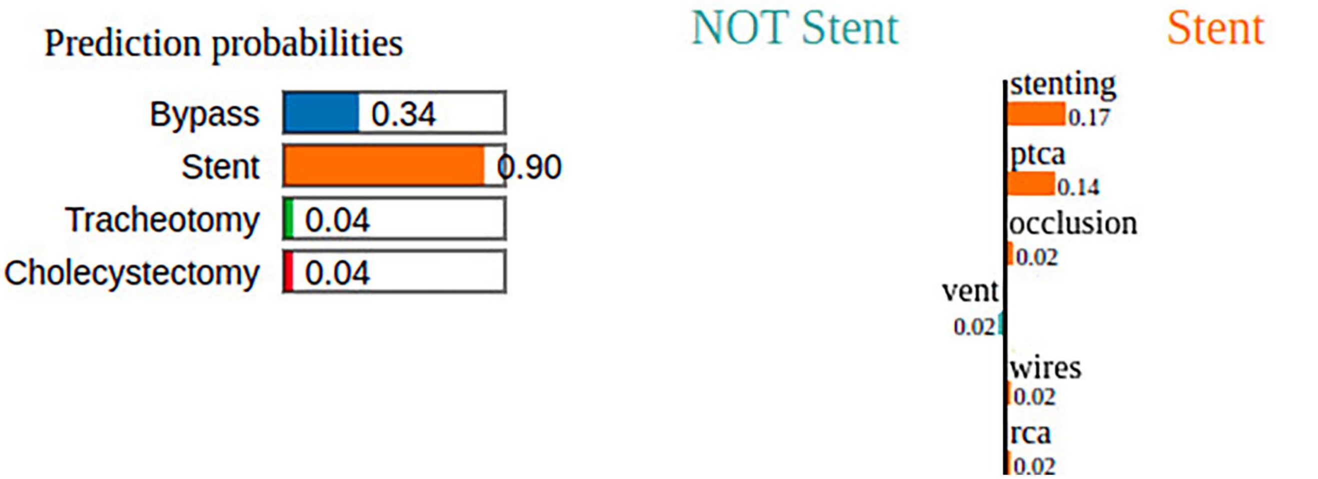


Figure 10. Example of an LIME explanation for prediction probabilities of our model for intervention prediction task where actual: "Stent" and predicted: "Stent." The bars' length highlights the specific contribution of each word of the nursing note: the orange ones push the model toward "Stent" prediction, whereas the other colors for other interventions. (A color version of this figure is available in the online journal.)

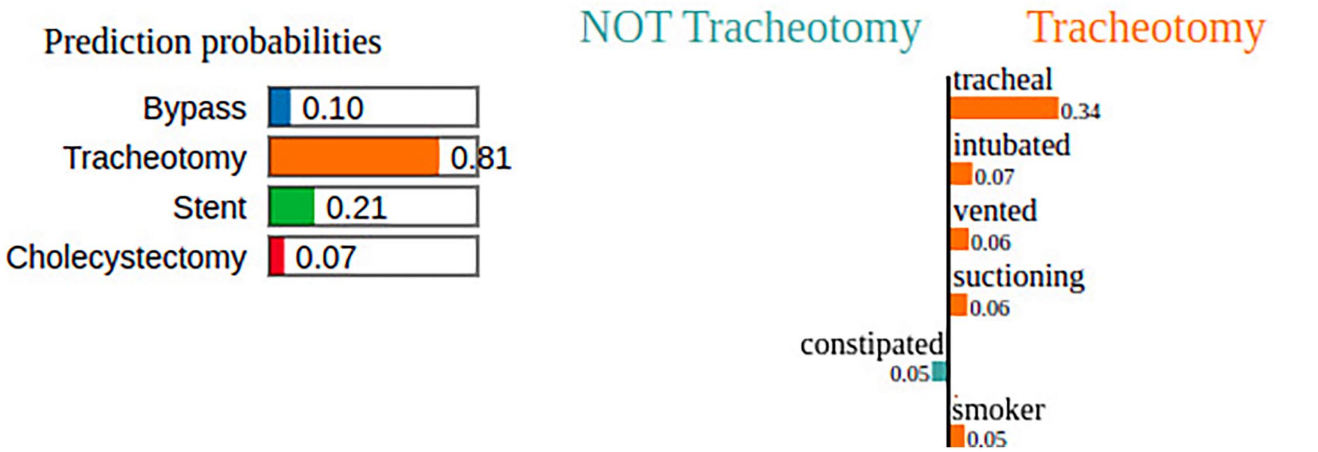


Figure 11. Example of an LIME explanation for prediction probabilities of our model for intervention prediction task where actual: "Tracheotomy" and predicted: "Tracheotomy." The bars' length highlights the specific contribution of each word of the nursing note: the orange ones push the model toward "Tracheotomy" prediction, whereas the other colors for other interventions. (A color version of this figure is available in the online journal.)

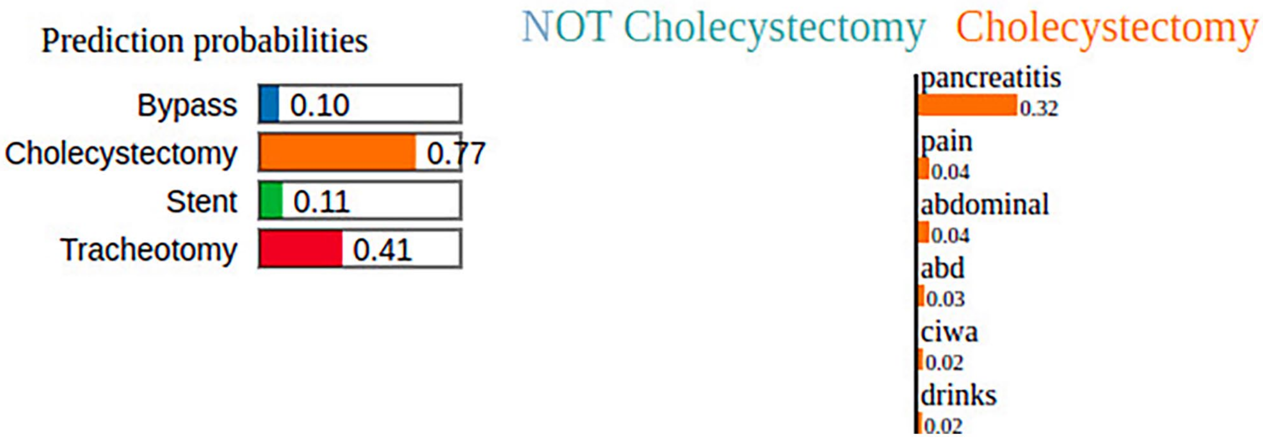


Figure 12. Example of an LIME explanation for prediction probabilities of our model for intervention prediction task where actual: "Cholecystectomy" and predicted: "Cholecystectomy." The bars' length highlights the specific contribution of each word of the nursing note: the orange ones push the model toward "Cholecystectomy" prediction, whereas the other colors for other interventions. (A color version of this figure is available in the online journal.)

"Tracheotomy," and "Cholecystectomy," respectively. Each of them depicts feature probability graphs that give insight into the extent of influence of the features on the outcome. In the prediction of "Bypass," LIME explainer emphasized the presence of features like "cardiac arrest," "CPAP," and "vent," while for the class "Stenting," the explainer gives importance to features like "PTCA" and "occlusion." For prediction of "Tracheotomy," the explainer gives higher weightage on the features "tracheal sectioning" and "intubated," while for "Cholecystectomy," the higher weighted features are "pancreatitis," "abdominal," and "pain."

In this study, we focused on using only the nursing notes generated within the first 24h of ICU admission for the prediction of ICU length of stay and a few interventions. While analyzing the performance of the model, we found that there exist many cases in the data set, for which the nursing notes do not contain sufficient information. Besides, these notes also contain non-standard abbreviations, medical syntax, and so on. The proposed architecture had difficulty in parsing these abbreviations and jargons, because of which performance suffered. To overcome this in the future, we plan to perform some domain-based preprocessing of the text to be used and also use more input resources from EHR clinical documentation like physician's notes, ECG reports, radiology reports, and so on for the predictions.

Conclusions

Predictive analytics is an important aspect of health-care informatics. Predicting the length of ICU stay and interventions have both been considered as fundamental problems in the domain of health-care informatics. Early predictions for ICU usage help in better management of hospital facilities. Predicting critical interventions for patients well in advance not only helps in efficient hospital resource management but also plays a significant role in saving patients' lives and contributes toward better expectation management for patients' families.

In this article, we have presented an explainable transformer-based multitask neural network architecture that predicts the length of stay of a patient in ICU and requirements for critical interventions based on the nursing notes recorded during the first 24h of admission. The model has been built using the publicly available MIMIC-III data set, which contains anonymized EHRs. Initial analysis of the data revealed that nursing notes, which are a part of the EHRs, contained a detailed description of a patient's condition, treatment plans, observations about the response to treatment, and many other relevant details in unstructured form. We built a multiobjective prediction model using this wealth of information from the nursing notes along with the clinical parameters. Our results demonstrate that the proposed model performs better than earlier models proposed for similar tasks. Although we have currently used the model for predicting one or more of four critical interventions, it can be easily adapted for predicting other interventions such as aortic valve replacement, pacemaker implant, and so on. We have utilized the LIME framework for generating explanations for the decisions. The LIME model can find the key features responsible for a particular prediction, and therefore serve as an important tool to provide explanations to the end-users. Explainability helps in gaining trust in the model. In the future study, we plan to utilize more textual information associated with patients' radiology reports, electroencephalogram (EEG) reports, physician notes, and so on, and also build multimodal models that can utilize image and sensor signals.

AUTHORS' CONTRIBUTIONS

All authors participated in the design, interpretation of the studies and analysis of the data, and review of the manuscript. SJ conducted the experiments. SJ, TD, and LD obtained the Institutional Review Board (IRB) approval and completed the training course "Data or Specimens Only Research" provided by the Collaborative Institutional Training Initiative (CITI) and got access to Medical Information Mart for Intensive Care (MIMIC) data. SJ drafted the manuscript, and TD and LD contributed to the manuscript preparation.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID ID

Sudeshna Jana https://orcid.org/0000-0003-4627-2987

REFERENCES

1. Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* 2010;**38**:65–71
2. Marshall JC, Bosco L, Adhikari NK, Connolly B, Diaz JV, Dorman T, Fowler RA, Meyfroidt G, Nakagawa S, Pelosi P, Vincent JL, Vollman K, Zimmerman J. What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. *J Crit Care* 2017;**37**:270–6
3. Chan CW, Farias VF, Escobar GJ. The impact of delays on service times in the intensive care unit. *Manage Sci* 2017;**63**:2049–72
4. Charles D, Gabriel M, Furukawa MF. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2012. *ONC Data Brief* 2013;**9**:1–9
5. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:160035
6. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019;**6**:96
7. Rocheteau E, Liò P, Hyland S. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: *Proceedings of the conference on health, inference, and learning*, 8–10 April 2021, pp.58–68. New York: ACM
8. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018;**5**:180178
9. Su L, Xu Z, Chang F, Ma Y, Liu S, Jiang H, Wang H, Li D, Chen H, Zhou X, Hong N, Zhu W, Long Y. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. *Front Med* 2021;**8**:664966
10. Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. *JMIR Med Inform* 2021;**9**:e21347
11. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. *J Biomed Inform* 2017;**69**:218–29
12. Van Aken B, Papaioannou J-M, Mayrdorfer M, Budde K, Gers FA, Löser A. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In: *Proceedings of the 16th conference of the European chapter of the association for computational linguistics*, 19–23 April 2021, pp.881–93. Stroudsburg, PA: Association for Computational Linguistics (ACL)
13. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. *BMC Med Inform Decis Mak* 2021;**21**:351
14. Huang K, Gray TF, Romero-Brufau S, Tulsky JA, Lindvall C. Using nursing notes to improve clinical outcome prediction in intensive care patients: a retrospective cohort study. *J Am Med Inform Assoc* 2021; **28**:1660–6
15. Mahbub M, Srinivasan S, Danciu I, Peluso A, Begoli E, Tamang S, Peterson GD. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLoS ONE* 2022;**17**:e0262182
16. Suresh H, Hunt N, Johnson AEW, Celi LA, Szolovits P, Ghassemi M. Clinical intervention prediction and understanding using deep networks. In: *Proceedings of the machine learning for healthcare (MLHC), Northeastern University*, Boston, MA, 18–19 August 2017
17. Bardak B, Tan M. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artif Intell Med* 2021;**117**:102112
18. Rumeng L, Abhyuday NJ, Hong Y. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. *AMIA Annu Symp Proc* 2017;**2017**:1149–58
19. Zhu X, Song B, Shi F, Chen Y, Hu R, Gan J, Zhang W, Li M, Wang L, Gao Y, Shan F, Shen D. Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. *Med Image Anal* 2021;**67**:101824
20. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proceedings of the 18th BioNLP workshop and shared task*, Florence, 1 August 2019, pp.58–65. Stroudsburg, PA: Association for Computational Linguistics (ACL)
21. Peng Y, Chen Q, Lu Z. An empirical study of multi-task learning on BERT for biomedical text mining. In: *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, 9 July 2020, pp.205–14. Stroudsburg, PA: Association for Computational Linguistics (ACL)
22. Lee L-H, Lu Y, Chen P-H, Lee P-L, Shyu K-K. NCUEE at MEDIQA 2019: medical text inference using ensemble BERT-BiLSTM-Attention model. In: *Proceedings of the 18th BioNLP workshop and shared task*, Florence, 1 August 2019, pp.528–32. Stroudsburg, PA: Association for Computational Linguistics (ACL)
23. Zhang Y, Gong L, Wang Y. An improved TF-IDF approach for text classification. *J Zhejiang Univ: Sci A* 2005;**6**:49–55
24. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985;**13**:818–29
25. Larvin M, McMahon MJ. APACHE-II score for assessment and monitoring of acute pancreatitis. *Lancet* 1989;**2**:201–5
26. Salluh JI, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care* 2014;**20**:557–65
27. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;**270**:2957–63
28. De Mendonca A, Vincent JL, Suter PM, Moreno R, Dearden NM, Antonelli M, Takala J, Sprung C, Cantraine F. Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. *Intensive Care Med* 2000;**26**:915–21
29. Moreno R, Vincent JL, Matos A, Mendonça A, Cantraine F, Thijs L, Takala J, Sprung C, Antonelli M, Bruining H, Willatts S. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. *Intensive Care Med* 1999;**25**:686–96
30. Johnson AEW, Kramer AA, Clifford GD. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Crit Care Med* 2013;**41**:1711–8
31. Chen Q, Zhang L, Ge S, He W, Zeng M. Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study. *PeerJ* 2019;**7**:e7083
32. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, CA, 13–17 August 2016, pp.1135–44. New York: ACM
33. Wilfong EM, Lovly CM, Gillaspie EA, Huang LC, Shyr Y, Casey JD, Rini BI, Semler MW. Severity of illness scores at presentation predict ICU admission and mortality in COVID-19. *J Emerg Crit Care Med* 2021;**5**:7
34. Stephens JR, Stümpfle R, Patel P, Brett S, Broomhead R, Baharlo B, Soni S. Analysis of critical care severity of illness scoring systems in patients with coronavirus disease 2019: a retrospective analysis of three U.K. ICUs. *Crit Care Med* 2021;**49**:e105–7
35. Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. *JAMA* 1994;**272**:1049–55
36. Alexander JH, Smith PK. Coronary-artery bypass grafting. *N Engl J Med* 2016;**374**:1954–64

37. Hannan EL, Racz MJ, Walford G, Jones RH, Ryan TJ, Bennett E, Culliford AT, Isom OW, Gold JP, Rose EA. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. *N Engl J Med* 2005;**352**:2174–83
38. Durbin CG. Tracheostomy: why, when, and how? *Respir Care* 2010;**55**: 1056–68
39. McSherry CK. Cholecystectomy: the gold standard. *Am J Surg* 1989; **158**:174–8
40. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the NAACL-HLT*, Minneapolis, MN, 2–7 June 2019, pp.4171–86. Stroudsburg, PA: Association for Computational Linguistics (ACL)
41. Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, Wallace BC. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. *Proc Conf Assoc Comput Linguist Meet* 2018;**2018**:197–207
42. Selva Birunda S, Kanniga Devi R. A review on word embedding techniques for text classification. In: Raj JS, Bashar A, Ramson SRJ (eds) *Innovative data communication technologies and application*. Berlin; Heidelberg: Springer, 2021, pp.267–281
43. Boag W, Wacome K, Naumann T, Rumshisky A. CliNER: a lightweight tool for clinical named entity recognition. In: *Proceedings of the AMIA joint summits on clinical research informatics (poster)*, 2015, https://knowledge.amia.org/amia-59309-cri2015-1.2002246/t-006-1.2003037/a-097-1.2003462/a-097-1.2003463/ap-097-1.2003464?timeStamp=1662594651262&qr=1

(Received May 5, 2022, Accepted August 15, 2022)