Original Paper

Artificial Intelligence–Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study

Weijia Chen[1*], MD; Zhijun Lu[1*], MD, PhD; Lijue You[2], MS; Lingling Zhou[3], BSc; Jie Xu[4,5], MPhil; Ken Chen[1,5,6], MD, DHM

[1]Department of Anesthesiology, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]Department of Informatics, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Infection Prevention and Control, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]VitalStrategic Research Institute, Shanghai, China

[5]Synyi Research, Shanghai, China

[6]Precision Diagnosis and Image Guided Therapy, Philips Research China, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Ken Chen, MD, DHM
Department of Anesthesiology
Rui Jin Hospital, Luwan Branch
Shanghai Jiao Tong University School of Medicine
South Chongqing Road, No 149
Shanghai
China
Phone: 86 021 63864050
Email: nutastray@gmail.com

Abstract

Background: Surgical site infection (SSI) is one of the most common types of health care–associated infections. It increases mortality, prolongs hospital length of stay, and raises health care costs. Many institutions developed risk assessment models for SSI to help surgeons preoperatively identify high-risk patients and guide clinical intervention. However, most of these models had low accuracies.

Objective: We aimed to provide a solution in the form of an Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection (AMRAMS) for inpatients undergoing operations, using routinely collected clinical data. We internally and externally validated the discriminations of the models, which combined various machine learning and natural language processing techniques, and compared them with the National Nosocomial Infections Surveillance (NNIS) risk index.

Methods: We retrieved inpatient records between January 1, 2014, and June 30, 2019, from the electronic medical record (EMR) system of Rui Jin Hospital, Luwan Branch, Shanghai, China. We used data from before July 1, 2018, as the development set for internal validation and the remaining data as the test set for external validation. We included patient demographics, preoperative lab results, and free-text preoperative notes as our features. We used word-embedding techniques to encode text information, and we trained the LASSO (least absolute shrinkage and selection operator) model, random forest model, gradient boosting decision tree (GBDT) model, convolutional neural network (CNN) model, and self-attention network model using the combined data. Surgeons manually scored the NNIS risk index values.

Results: For internal bootstrapping validation, CNN yielded the highest mean area under the receiver operating characteristic curve (AUROC) of 0.889 (95% CI 0.886-0.892), and the paired-sample *t* test revealed statistically significant advantages as compared with other models ($P<.001$). The self-attention network yielded the second-highest mean AUROC of 0.882 (95% CI 0.878-0.886), but the AUROC was only numerically higher than the AUROC of the third-best model, GBDT with text embeddings (mean AUROC 0.881, 95% CI 0.878-0.884, $P=.47$). The AUROCs of LASSO, random forest, and GBDT models using text embeddings were statistically higher than the AUROCs of models not using text embeddings ($P<.001$). For external validation, the self-attention network yielded the highest AUROC of 0.879. CNN was the second-best model (AUROC 0.878), and GBDT

with text embeddings was the third-best model (AUROC 0.872). The NNIS risk index scored by surgeons had an AUROC of 0.651.

Conclusions: Our AMRAMS based on EMR data and deep learning methods—CNN and self-attention network—had significant advantages in terms of accuracy compared with other conventional machine learning methods and the NNIS risk index. Moreover, the semantic embeddings of preoperative notes improved the model performance further. Our models could replace the NNIS risk index to provide personalized guidance for the preoperative intervention of SSIs. Through this case, we offered an easy-to-implement solution for building multimodal RAMs for other similar scenarios.

(JMIR Med Inform 2020;8(6):e18186) doi: 10.2196/18186

KEYWORDS

surgical site infection; machine learning; deep learning; natural language processing; artificial intelligence; risk assessment model; routinely collected data; electronic medical record; neural network; word embedding

Introduction

Health care–associated infection (HAI) is a global patient safety problem, with surgical site infection (SSI) being one of the most common types of HAI [1-4]. The incidences of SSI for inpatients undergoing operations are 2%-5% in the United States [5], 2%-10% in Europe [6-9], and 4%-6% in China [10-13]. SSIs increase mortality and long-term disabilities, prolong hospital length of stay (LOS), and raise health care costs [1,5,11]. In China, SSIs prolong hospital LOS by 6-23 days and increase medical costs by US $2000-$6000 per patient, with the additional cost for one SSI patient needing to be offset by the medical revenue from 13 surgical patients [11].

In 2016, the World Health Organization recommended a large perioperative care bundle of interventions for preventing SSIs, which includes perioperative oxygen inhalation; maintenance of normal body temperature; maintenance of adequate glucose and circulating volume; use of sterile drapes, surgical gowns, wound-protector devices, and antimicrobial-coated sutures; provision of incisional wound irrigation; and prophylactic negative-pressure wound therapy [14]. However, the quality of evidence for most of these recommended interventions remains low. When we do not know whether these interventions are effective enough, using several interventions together is reasonable, and may even have a summation effect, for reducing the risk of SSI as much as possible. However, the shortcomings of bundle interventions are also apparent: they will consume large amounts of medical resources, especially when we strictly implement the recommendations of the guideline. Thus, data-driven guidance for personalized intervention is key to creating more effective SSI prevention and control programs.

Many institutions have developed risk assessment models (RAMs) focusing on SSIs to help surgeons preoperatively identify high-risk patients and guide clinical interventions. The most widely used traditional RAM is the National Nosocomial Infections Surveillance (NNIS) risk index [15], which is a scoring system ranging from 0 to 3. An American Society of Anesthesiologists (ASA) preoperative assessment score higher than 2; contaminated, dirty, or infected operation; and prolonged operation duration each account for 1 point in the NNIS risk index scoring system. The risk of SSI increases from 1.5% to 13.0% as the score goes up. Obviously, the three variables are easy to calculate, but are not enough to describe the characteristics of high-risk patients. To remedy these deficiencies, Mu et al included more patient- and hospital-specific variables and developed improved RAMs for each procedure under the 39 National Healthcare Safety Network (NHSN) procedure categories using stepwise logistic regression [16]. They trained these procedure-specific models using 849,659 patient records, from 2006 to 2008, from the NHSN database. Each model used 12-15 variables, including patient demographics, anesthesia, surgery, hospital settings, and NNIS risk index factors. The overall area under the receiver operating characteristic curve (AUROC) of the model reached 0.67, higher than the AUROC of the NNIS risk index, which is 0.60. The biggest problem with their RAM is that 39 different models need to be deployed together to achieve full functionality, which is cumbersome for clinical use. In a later study, Grant et al developed another RAM, using routinely collected surveillance data from three national networks in Switzerland, France, and England [17]. They trained a logistic regression model using 46,320 colorectal surgery records from 2007 to 2017 and compared it with the previous model developed by Mu et al. In their dataset, the new model, with an AUROC of 0.65, outperformed the model developed by Mu et al. Their model was easy to use but was limited to colorectal surgery only. Meanwhile, in the absence of a high-accuracy RAM, van Walraven and Musselman developed a logistic regression model based on 362,431 clinical data points from the National Surgical Quality Improvement Program [18]. The AUROC of this model reached 0.80. However, it required the users to provide large amounts of medical history information, such as ASA score, NNIS risk index, tumor history, medication history, and operation history. These variables are not always well structured in many electronic medical record (EMR) systems, and without the support of automatic extraction, completing evaluations based on this model undoubtedly consumed large amounts of time. Therefore, a gap still exists between current preoperative RAMs and the ideal RAM, which is generalized, accurate, and easy to use or deploy.

With the widespread use of hospital information systems and EMR systems in medical institutions, we can now use massive clinical data to build RAMs. In addition to structured data, we can also use natural language processing and deep learning technology to parse semantics from unstructured clinical text data and save time for manual extraction of text information. Many researchers have developed surveillance models using data from EMRs to automatically help infection control staff

efficiently identify SSIs among massive medical records and have achieved high accuracy [19-22]. However, these models used not only preoperative information but also surgical, postoperative, and antibiotic information. Thus, they cannot be used to guide preoperative intervention.

To fill in the gap, we aimed to provide a solution in the form of the Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection (AMRAMS) for inpatients undergoing operations using routinely collected data from the EMR system of a general hospital in China. We believed that structured data, such as patient demographics and preoperative lab results, and free-text data, such as preoperative notes that record diagnoses and scheduled surgical information, would both help to identify high-risk patients. Thus, we planned to combine various machine learning, deep learning, and semantic representation technologies; validate the discriminations of multimodal implementations internally and externally; and compare them with the NNIS risk index score. We tested the following hypotheses: (1) AMRAMS, with various implementations, would more accurately identify high-risk patients than the old-fashioned NNIS risk index is capable of doing, (2) semantic information from preoperative notes would improve the model performance, and (3) deep learning implementations would outperform conventional machine learning implementations.

Methods

Source of Data

The Rui Jin Hospital, Luwan Branch, affiliated with the Shanghai Jiao Tong University School of Medicine, is a nonprofit academic medical center based in Huangpu District, Shanghai, China. The hospital has a total of 526 beds, of which 89 are in general surgery, 33 are in gynecology, 27 are in orthopedics, and 38 are in urology. The surgical staff performs more than 4000 operations annually. About 300 of these cases are emergency patients. We retrieved inpatient records that each had only one operation record during the hospital stay and a discharge record between January 1, 2014, and June 30, 2019, from the EMR system of Rui Jin Hospital, Luwan Branch. We used data from before July 1, 2018, as the development set for model training, hyperparameter tuning, and internal validation; we used the remaining data as the test set for external validation. The data usage of patient records for this study had been reviewed and approved by the ethics committee of the Rui Jin Hospital, Luwan Branch.

Participants and Features

We included adult patients only and excluded patients under the age of 18 years, patients with missing operation information (ie, timestamp of operation, whether or not theirs was an emergency operation, and type of anesthesia), and patients with missing demographic information (ie, gender and age).

We used both structured and unstructured preoperative clinical data from the EMR as our modeling features for this study. *Preoperative* was defined as the last record before the timestamp of the operation start time. Structured data included the following:

1. Patient demographics: age (years), gender (male or female), body height (cm), body weight (kg), and type of insurance (insured or noninsured).
2. Routine blood examination: white blood cell count (number $\times 10^9$/L), proportion of neutrophils (%), proportion of lymphocytes (%), proportion of monocytes (%), proportion of eosinophils (%), proportion of basophils (%), lymphocyte count (number $\times 10^9$/L), monocyte count (number $\times 10^9$/L), eosinophil count (number $\times 10^9$/L), red blood cell count (number $\times 10^{12}$/L), hemoglobin concentration (g/L), mean corpuscular volume (fL), mean corpuscular hemoglobin (g/L), mean corpuscular hemoglobin concentration (g/L), and platelet count (number $\times 10^9$/L).
3. Coagulation function examination: prothrombin time (sec), international normalized ratio, fibrinogen concentration (g/L), activated partial thromboplastin time (sec), thrombin time (sec), and d-dimer concentration (mg/L).
4. Liver and kidney function examination: total bilirubin concentration (μmol/L), direct bilirubin concentration (μmol/L), indirect bilirubin concentration (μmol/L), total bile acid concentration (μmol/L), alanine transaminase concentration (IU/L), aspartate aminotransferase concentration (IU/L), total protein concentration (g/L), albumin concentration (g/L), urea nitrogen concentration (mmol/L), creatinine concentration (μmol/L), uric acid concentration (μmol/L), and blood glucose concentration (mmol/L).
5. Plasmic electrolyte examination: potassium concentration (mmol/L), sodium concentration (mmol/L), calcium concentration (mmol/L), phosphorus concentration (mmol/L), and magnesium concentration (mmol/L).
6. Structured data elements from admission notes: current smoking status (true or false) and marital history (married, unmarried, or divorced).
7. Structured data elements from preoperative notes: emergency operation (true or false) and type of anesthesia (general anesthesia, total intravenous anesthesia, spinal anesthesia, epidural anesthesia, nerve block, or local anesthesia).
8. Preoperative LOS: the number of inpatient days between admission and operation.

Unstructured data included the free-text portion of the preoperative notes. Preoperative notes usually contain descriptions about preoperative diagnosis, operation name, indication, complications, and preventive measures. Table MA1-1 in Multimedia Appendix 1 shows two examples of preoperative notes from the development set.

Outcome

According to the No. 48 Decree issued by the Ministry of Health of the People's Republic of China in 2006 [23], the infection prevention and control department of the hospital should be responsible for the regular surveillance, analysis, and feedback on the epidemic situation of SSIs and their related risk factors. SSIs includes superficial incisional infection, deep incisional infection, and organ-space infection. In the Rui Jin Hospital, Luwan Branch, the staff of the infection prevention and control department manually identify SSIs via patient chart reviews

and collect mandatory data for hospital administrators and government reporting after patient discharge. In this study, all the included patient records were reviewed by the infection prevention and control department. We categorized patient records with SSI identifications reported by the infection prevention and control department as positive samples. Likewise, we categorized patient records without SSI identifications as negative samples.

Data Preprocessing

In this study, we used routinely collected clinical data from the EMR system. Thus, outliers and missing data were common and inevitable. For outlier adjustment, we first discarded patient records with invalid age values (eg, age >120 years old). To detect the outliers, we implemented a two-stage algorithm based on the random-effects model adjusted for age and sex proposed by Welch et al [24] for all continuous features. We used an absolute standardized residual of more than 5 as a cutoff for outlier detection and manually reviewed all the outliers suggested by the model. If the outliers violated medical knowledge, we tried to correct the values via a chart review. If no information could be gained from the chart review, we considered the outliers as missing values. For missing data, we generated missing-value indicators [25] (ie, binary dummy features indicating whether the values of the original features were missing) and conducted mean imputation for continuous features and mode imputation for binary or categorical features. To make the results of model validation truly reflect real-world performance, we performed outlier detection and adjustment only on the development set. For the convenience of model training and optimization, we performed one-hot encoding for all the binary or categorical features and feature normalization for all features.

Model

Overview

Using the fastText algorithm, we first generated word embeddings based on a large Chinese corpus for further text-information encoding [26]. The fastText algorithm is an unsupervised neural network algorithm that learns distributional embeddings of semantic representation based on subword information for each word from the corpus. We then proposed both conventional machine learning methods and deep learning methods to predict the risk of SSI based on preoperative EMR data. Because we expected the distribution of the labels to be extremely imbalanced, we passed a positive sample weight (ie, the ratio of the number of non-SSIs to the number of SSIs) to the loss function (ie, cross-entropy) during the model training. We used the mini-batch gradient descent and backpropagation technique to update the parameters of the networks and set the AdamW algorithm as optimizer [27]. Furthermore, we used a random search method based on five-fold cross-validation and early stopping, if necessary, to find the optimized hyperparameters for each model.

Word Embedding

Our Chinese corpus contained approximately 4.1 GB of data from the Chinese Wikipedia, downloaded from the linguatools website [28], and approximately 96.9 MB of data from A-hospital, a Chinese medical Wikipedia website [29]. After we removed punctuations and numbers from the corpus, we used Jieba, version 0.41 [30], a Chinese text-segmentation tool, with a medical dictionary to segment the corpus into word sequences. Using the skipgram model of the fastText algorithm [31], we then trained 128-dimensional word embeddings using the preprocessed word sequences. We set the minimum size of the subwords as two and the maximum size as five. We left the other parameters at their default values.

Conventional Machine Learning Method

The conventional machine learning methods analyzed in this study included LASSO (least absolute shrinkage and selection operator) logistic regression with L1 penalty, random forest, and gradient boosting decision tree (GBDT), implemented by the XGBoost framework [32]. Because these models can be trained only by using tabular data, we first encoded the texts of the preoperative notes into the text embeddings. We segmented each text into a word sequence using Jieba and transformed it into a sequence of word embeddings, which is represented as follows:

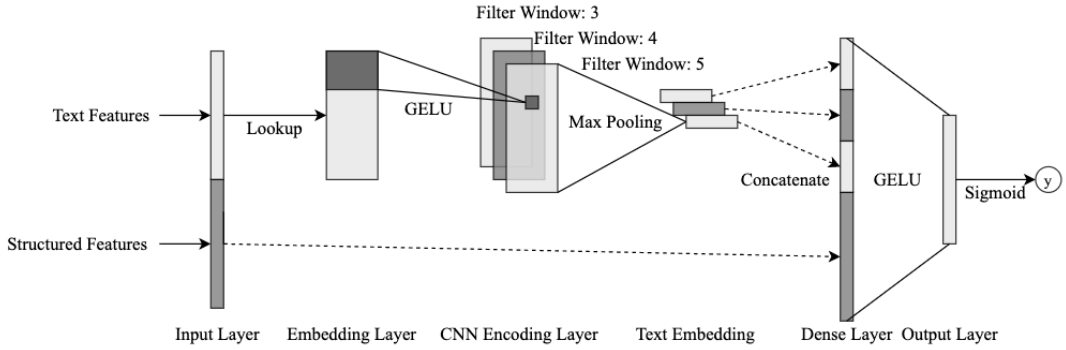
$$T = [t_1,t_2,t_3,...,t_n] \quad (1)$$

Here, t_i is a 128-dimensional word-embedding vector of the i-th word in a sequence of length n, and T is an n-by-128 embedding matrix. We then pooled the embedding matrix into a 128-dimensional vector using the max-pooling method, which took the maximum value among the n words for each feature of the word embeddings. We concatenated the pooled vectors with the structured feature vectors and fed them into the four models for training.

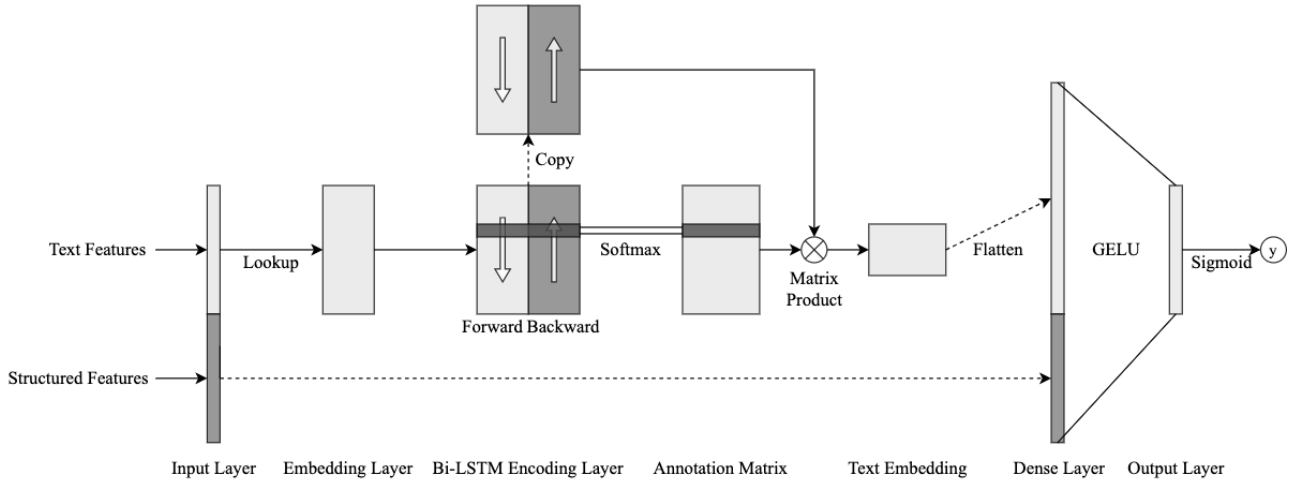
Deep Learning Methods

The deep learning methods analyzed in this study included a convolutional neural network (CNN) and a self-attention network. In this study, we used CNN and self-attention structures to encode text information on an end-to-end basis. Figure 1 shows the architecture of both models. Before being fed into the models, text data were transformed into n-by-128 embedding matrix T. Here, n was the padding length decided by the upper boundary of the 1.5-IQR rule based on the distribution of word sequence lengths in the development set. If the actual length of the sequence was less than n, we added zero-padding to the left side of the sequence. If the actual length was more than n, we tailored the left side of the sequence to suit the padding length.

Figure 1. Deep learning network architecture diagrams. Bi-LSTM: bidirectional long short-term memory; GELU: Gaussian Error Linear Unit.



(A) Architecture Diagram of Convolutional Neural Network (CNN)



(B) Architecture Diagram of Self-attention Network

For the CNN, we referred to the structure proposed by Kim [33] and applied convolutional kernel operation on the embedding matrix T, which is represented as follows:

$$c^j_i = GELU(w^j_c \cdot t_{i:I+h-1} + b) \quad (2)$$

Here, c^j_i is the output value of the j-th convolutional channel of filter window i, $t_{i:i+h-1}$ is the word-embedding sequence from the i-th word to the $(i + h - 1)$-th word, h is the size of the filter window, w^j_c is the weight vector for each word embedding in the filter window of the j-th channel, b is the bias item, and *GELU* (Gaussian Error Linear Unit) [34] is the active function; the original paper used *ReLU* (Rectified Linear Unit). The filter window slides from the first word to the last one. Let m represent the number of channels for the convolutional kernel, and let n represent the length of the text; then we have an $(n - h + 1)$-by-m convolutional feature matrix:

$$C = [c_1,c_2,c_3,...,c_{n-h+1}] \quad (3)$$

We used filter windows of three, four, and five; generated three convolutional feature matrices; applied max-pooling operation on these matrices; and concatenated the three pooled vectors and the structured feature vector together. The entire vector was then passed into a fully connected dense layer using *GELU* as an active function and, finally, a *sigmoid* output layer.

For the self-attention network, we referred to the structure proposed by Lin et al [35]. The embedding matrix T was first passed into a bidirectional long short-term memory (Bi-LSTM) layer. Let m represent the number of the hidden units for both forward and backward long short-term memory (LSTM), and let n represent the length of the word sequence. We obtained an n-by-$2m$ hidden state feature matrix:

$$H = [h_1,h_2,h_3,...,h_n] \quad (4)$$

We then generated an attention matrix based on the hidden state feature matrix. According to the original paper, this process was similar to passing the hidden state feature matrix into two bias-free, fully connected layers using *tanh* as the first active function and *softmax* as the second function:

$$A = softmax(W_1 \cdot tanh(W_2 \cdot H^T)) \quad (5)$$

Here, W_2 is a d_1-by-$2m$ weight matrix, where d_1 is the hidden unit number of the first layer and W_2 is a d_2-by-d_1 weight matrix, where d_2 is the hidden unit number of the second layer. We obtained a d_1-by-$2m$ text-embedding matrix by using the following:

$$M = A \cdot H \quad (6)$$

We flattened the embedding matrix into a $d_1 \times 2m$-dimensional vector, concatenated it with the structured feature vector, and passed the entire vector into the dense layer.

Evaluation

We evaluated the discrimination capacities of the conventional models, with or without text embeddings, and of the deep learning models. We assessed internal validity using a bootstrapping procedure of 100 iterations based on the development set. In each iteration, we trained the model using the sample data (ie, sampling with replacement) and tested the AUROC on the out-of-bag data; the data were not sampled in the iteration. We calculated the average AUROC and 95% CI for each model and performed paired-sample *t* tests to compare the performance among the models.

To obtain a realistic estimation of the model performance, we assessed external validity on the holdout test set. We trained the models using the entire development set, tested the full training performances on the test set, and compared them with the performance of the NNIS risk index scored by surgeons before an operation. Furthermore, we calculated the sensitivity and specificity based on the test result and decided on the cutoff point using Youden's index method [36].

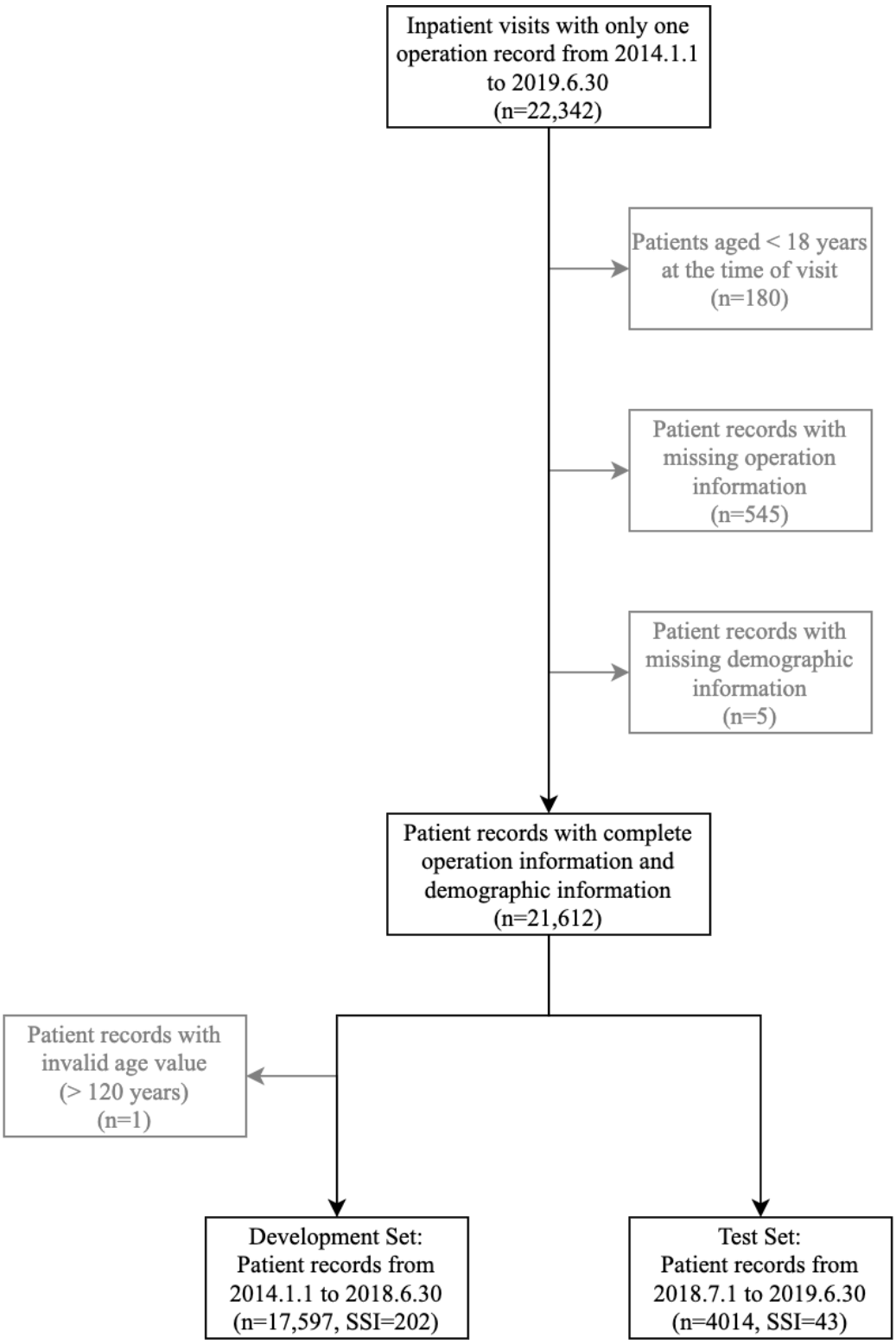
We used scikit-learn, version 0.22.1, via Python, version 3.7.4 (Python Software Foundation), to build the LASSO model and random forest model; we used XGBoost, version 0.9, via Python to build the GBDT model; and we used PyTorch, version 1.4.0, via Python to build the CNN and self-attention network. We performed all the statistical analyses using R, version 3.6.1 (The R Foundation), and considered a two-sided *P* value of <.05 as statistically significant.

Results

Patient Characteristics

We included a total of 21,611 inpatient records from January 1, 2014, to June 30, 2019. Of these records, 13,293 (61.51%) were from female patients and 8318 (38.49%) were from male patients with a median age of 54.3 years (IQR 44-65); 8375 (38.75%) were from the department of general surgery; 5903 (27.31%) were from the department of urology; 4649 (21.51%) were from the department of gynecology; and 2684 (12.42%) were from the department of orthopedics. According to the distributions of the International Classification of Diseases, Tenth Revision (ICD-10) code of operation and diagnosis that were retrieved after patient discharge, the patients received surgical treatment mainly for genitourinary system diseases, neoplasms, and digestive system diseases; the main types of operations were urinary system surgery, digestive system surgery, female reproductive system surgery, and endocrine system surgery. Overall, the incidence of SSIs in our dataset was 1.13% (244/21,611). The assigned sample size of the development set was 17,597 and that of the test set was 4014. The missing-data rates of the included variables ranged from 0% to 70.9% in the development set and from 0% to 72.8% in the test set. The variables with missing-data rates of more than 20% came from liver and kidney function examination, plasmic electrolyte examination, and d-dimer measurement. A slight difference was observed in the missing-data rate of each variable between the development set and the test set. Among them, the variables with the largest differences in the missing-data rates came from the electrolyte examinations (ie, calcium, phosphorous, and magnesium), with the rate differences reaching 8.0%. Figure 2 shows the selection process for the patient records and Table MA1-2 in Multimedia Appendix 1 shows the patient characteristics. We released a portion of the raw data in Multimedia Appendix 2, and the data dictionary of the raw data is located in the Data Description section of Multimedia Appendix 1.

Figure 2. Flowchart of the selection process for patient records. Gray boxes show records that were excluded due to patients not meeting inclusion criteria and records containing outliers or missing data. SSI: surgical site infection.



Hyperparameters and Training

We selected the optimal hyperparameters for each model based on the results of five-fold cross-validation. For LASSO, we used an L1 penalty of 0.01 when using text embeddings and 0.003 when not using text embeddings. For random forest with text embeddings, we used 300 trees, a maximum depth of 18, and maximum features of 0.6. For random forest without text embeddings, we used 1000 trees, a maximum tree depth of 4, and maximum features of 0.6. For GBDT with text embeddings, we used a learning rate (η) of 0.01, a maximum tree depth of 24, a subsample of 0.6, a column sample of 0.65, a gamma of 0.3, and 61 iterations. For GBDT without text embeddings, we used a learning rate (η) of 0.003, a maximum tree depth of 4, a subsample of 0.65, a column sample of 0.8, a gamma of 0, and 132 iterations. For the CNN, we used a learning rate (η) of

0.0001; an L2 penalty of 3; a word-embedding layer dropout rate of 0; CNN filter windows of three, four, and five with 256 feature maps (ie, channels) each; a dropout rate of 0.35; a fully connected layer with 128 feature maps; a dropout rate of 0.5; and 18 epochs. For the self-attention network, we used a learning rate (η) of 0.0001, an L2 penalty of 0.03, a word-embedding layer dropout rate of 0.5, a Bi-LSTM with 256 feature maps (ie, hidden nodes) each, a dropout rate of 0.45, an attention network with 256 feature maps (ie, hidden nodes) on the first layer and 64 for the second layer, a fully connected layer with 128 feature maps, a dropout rate of 0, and 19 epochs. We set the padding length for deep learning to 244. Hyperparameters not mentioned in this section were left at their default values.

Model Performances

Table 1 lists the performances of the models in terms of both internal and external validation, and Figure 3 shows the receiver operating characteristic (ROC) curves of the top five models based on full training and NNIS risk index. For internal validation, CNN yielded the highest mean AUROC of 0.889 (95% CI 0.886-0.892), and the paired-sample *t* test (see Multimedia Appendix 1, Table MA1-3) revealed statistically significant advantages ($P<.001$) compared with the other models. The self-attention network yielded the second-highest mean AUROC of 0.882 (95% CI 0.878-0.886). However, the AUROC of the self-attention network was only numerically higher than the AUROC of the third-best model—GBDT with text embeddings (mean AUROC 0.881, 95% CI 0.878-0.884)—and did not exhibit statistical significance ($P=.47$). The AUROCs of the machine learning models using text embeddings were statistically higher than the AUROCs of the models not using text embeddings ($P<.001$). For external validation, the self-attention network yielded the highest AUROC of 0.879. CNN was the second-best model (AUROC 0.878), and GBDT with text embeddings was the third-best model (AUROC 0.872). The NNIS risk index scored by surgeons had an AUROC of 0.651, which was remarkably lower than that of any other model in our study. Based on the external validation, we could still observe a trend with the text embeddings improving the model performances in the external validation. All the models had lower AUROC scores in internal validation than in external validation (ie, mean AUROC).

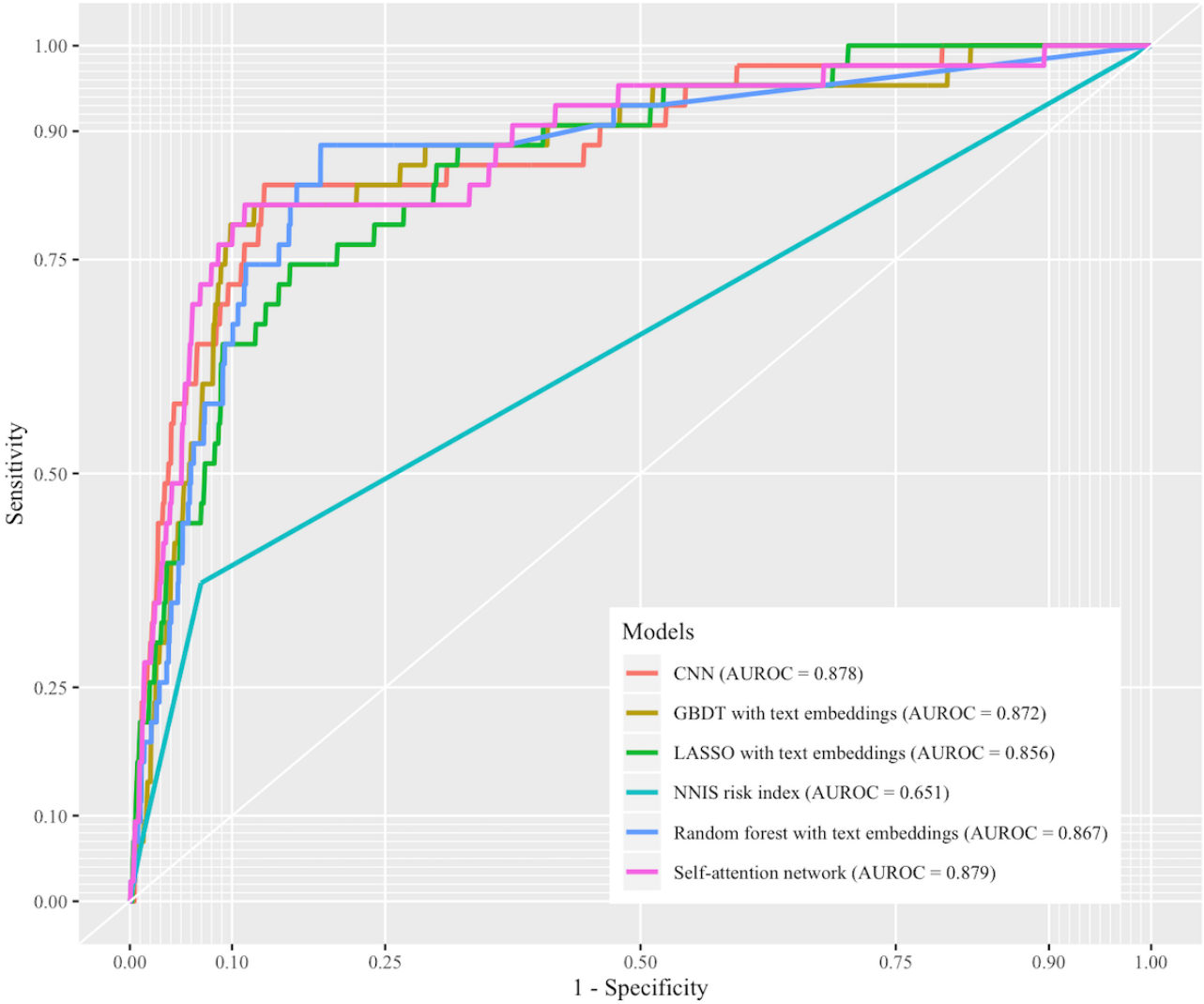
Table 1. Model performances.

Model and text embedding	Area under the receiver operating characteristic curve (AUROC)		Sensitivity[a] (full training)	Specificity[a] (full training)
	Bootstrapping, mean (95% CI)	Full training		
Least absolute shrinkage and selection operator (LASSO)				
With text embedding	0.870 (0.867-0.874)	0.856	0.744	0.844
Without text embedding	0.856 (0.852-0.860)	0.816	0.674	0.842
Random forest				
With text embedding	0.877 (0.873-0.880)	0.867	0.884	0.813
Without text embedding	0.846 (0.842-0.850)	0.772	0.558	0.871
Gradient boosting decision tree (GBDT)				
With text embedding	0.881 (0.878-0.884)	0.872	0.791	0.902
Without text embedding	0.838 (0.834-0.843)	0.782	0.605	0.858
Convolutional neural network (CNN)	0.889 (0.886-0.892)	0.878	0.837	0.869
Self-attention	0.882 (0.878-0.886)	0.879	0.814	0.888
National Nosocomial Infections Surveillance (NNIS) risk index	N/A[b]	0.651	0.372	0.930

[a]The optimal cutoff point was identified using Youden's index method.

[b]N/A: not applicable.

Figure 3. The receiver operating characteristic (ROC) curves of the top five models based on full training and National Nosocomial Infections Surveillance (NNIS) risk index. AUROC: area under the receiver operating characteristic curve; CNN: convolutional neural network; GBDT: gradient boosting decision tree; LASSO: least absolute shrinkage and selection operator.



Feature Analysis

Both deep learning models—CNN and self-attention network—performed better than other models in our validations. However, the deep learning models were black boxes and hard to explain. To further explore the correlations between the selected features and the occurrence of SSIs, we conducted a population-level feature analysis for the structured features and a case-level analysis for the text embeddings. The population-level analysis explored the correlations by comparing the normalized coefficient for each feature from LASSO without text embeddings; the coefficients were based on the data after normalization. For case-level analysis, we referenced the idea from local interpretable model-agnostic explanations [37]. For each case, we fixed the structured features and generated new word sequences by randomly removing words from the raw sequence and dummy binary vectors that indicated whether the word in a certain position was removed or not. We generated 10,000 new sequences for each case, combined them with the structured features, passed them to the deep learning models, and obtained prediction scores. We then fitted a LASSO regression model—with an L1 penalty of 0.01—that uses dummy binary vectors as features and prediction scores as targets. The coefficients of the LASSO regression model indicated the relative contributions of the words to the prediction scores in a case.

Figure 4 shows the features with nonzero coefficients and their coefficients for the population-level analysis. We could observe that preoperative LOS, marital history, anesthesia type, gender, age, results of routine blood examination, coagulation function examination, and many missing-value indicators had remarkable impacts on the model. Among them, patients with prolonged preoperative LOS, patients with missing AST results, married patients, older patients, and patients with missing body weight information had higher risks of SSI. Patients with higher hemoglobin, female patients, patients with missing magnesium results, patients that had received total intravenous anesthesia, and patients with missing marital histories had lower risks of SSI.

Figure 4. The normalized coefficients of the features in the LASSO (least absolute shrinkage and selection operator) model without text embeddings. ALB: albumin; APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; CA: calcium; DBIL: direct bilirubin; EA: epidural anesthesia; GLU: blood glucose; HGB: hemoglobin; INR: international normalized ratio; K: potassium; LOS: length of stay; LYMPH: lymphocyte; MCH: mean corpuscular hemoglobin; MCV: mean corpuscular volume; MG: magnesium; MONO: monocyte; NA: sodium; PP: phosphorus; SA: spinal anesthesia; TBIL: total bilirubin; TIVA: total intravenous anesthesia; TP: total protein; TT: thrombin time; UA: uric acid.

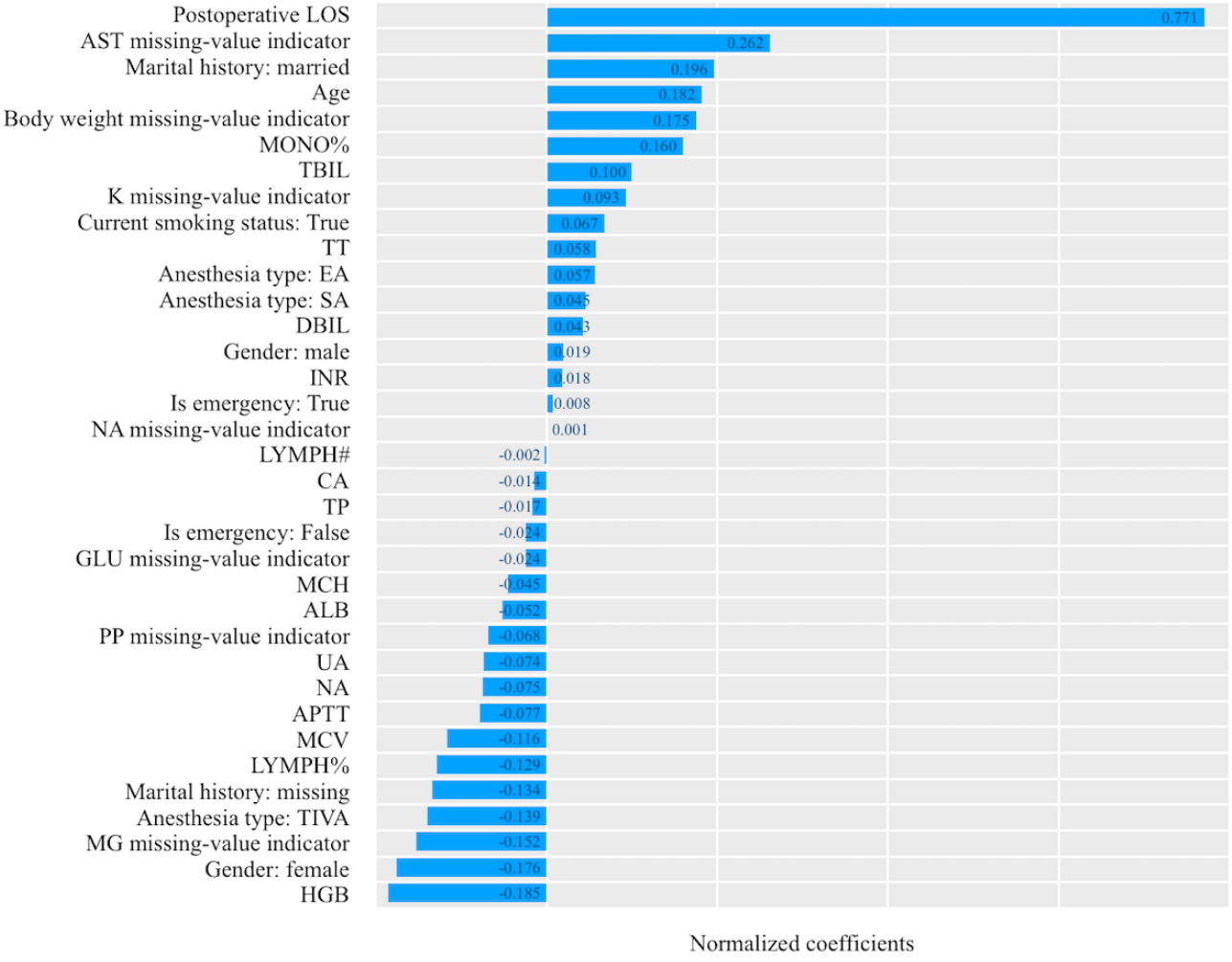
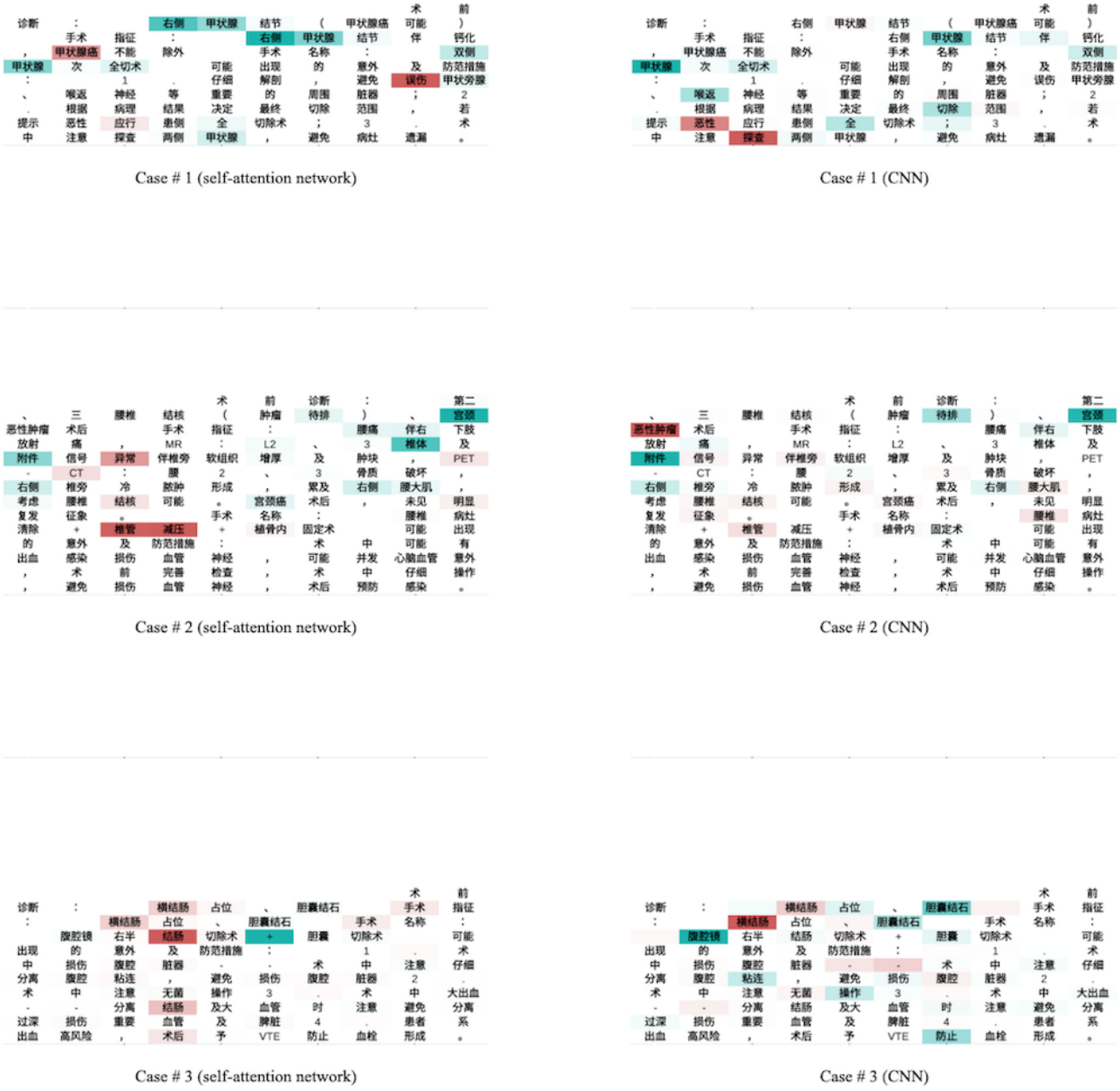


Figure 5 shows the heatmaps of the word contributions to the CNN and attention prediction scores for three preoperative note cases (see Table MA1-1 in Multimedia Appendix 1 for the full-text translation), with green being a negative coefficient (ie, protective factor) and red being a positive coefficient (ie, risk factor). The deeper the color, the higher the absolute value. Among the three cases, we could observe that terms like “甲状腺 (thyroid),” “宫颈 (uterine neck),” “附件 (accessory),” “椎体 (centrum),” “腹腔镜 (laparoscope),” and “胆囊结石 (gallstone)” were associated with lower risk of SSI, and terms like “甲状腺癌 (thyroid cancer),” “恶性 (malignant),” “结核 (tuberculosis),” “恶性肿瘤 (malignant tumor),” “结肠 (colon),” and “横结肠 (transverse colon)” were associated with higher risk of SSI.

Figure 5. The heatmaps of the word contributions on three preoperative note cases. CNN: convolutional neural network.



Discussion

Principal Findings

In this study, we found that SSI RAMs based on clinical data from an EMR system and modern machine learning techniques could identify high-risk patients more accurately than the old-fashioned NNIS risk index is capable of doing. Notably, the vectorial embedding of preoperative notes, whether generated using a simple max-pooling method or using a deep learning method, improved the performance of the model further without any handcrafted feature engineering. The multimodal deep learning models that produced end-to-end feature representations automatically through convolutional kernels, LSTM, or attention mechanisms outperformed the traditional machine learning models, such as LASSO, random forest, or GBDT. Thus, our AMRAMS using a CNN or a self-attention network could replace the NNIS risk index in providing personalized guidance for the preoperative intervention of SSI. At the same time, our study provided an easy-to-implement solution to building a multimodal RAM for similar scenarios based on both structured and Chinese text data. Because we used routinely collected preoperative data only, such as the results of routine blood tests and clinical notes, additional manual data collection and clinician evaluation was no longer necessary for achieving high accuracy.

Many factors could explain the advantages of the deep learning AMRAMS. First, we used more objective quantitative features of the patients. The NNIS risk index contained only three elements, of which the ASA score was a subjective feature provided by anesthesiologists. The model developed by Mu's team, on the other hand, utilized 12-15 features and included the ASA score [16]. The model developed by Grant's team utilized seven features and also included the ASA score [17]. The model developed by van Walraven and Musselman utilized

53 features and included not only subjective features, such as the ASA score, the NNIS risk index score, and dyspnea evaluation, but also 33 manually extracted variables from the medical history of the patient [18]. Meanwhile, our model included 47 objective features, among which age, gender, body weight, anesthesia type, emergency operation, preoperative hemoglobin, glucose, and LOS have proven to be related to the occurrences of SSI [12,38,39]. All these features were the results of routine preoperative blood tests and were automatically extracted from the EMR system using SQL (Structured Query Language) query script.

Second, we used sufficient information about the operative procedures and risk prevention from the preoperative notes via the fastText embeddings and the network structures. Many studies on English text classification have demonstrated that machine learning using fastText embeddings had better performance than those of the algorithms using bag-of-words, n-grams, TF-IDF (term frequency–inverse document frequency), or word2vec embeddings [26,40]. The semantics of many words, especially in the Chinese language, depend on the subwords or characters they contain. The fastText algorithm generated the semantic embeddings based on the internal structures of words, which best suit the characteristics of natural language. Moreover, our deep learning models enabled end-to-end learning: both fastText embeddings and hidden nodes of the network could be further fine-tuned simultaneously according to the specified targets during the learning process. To encode text-level semantics, we tried both convolutional kernel and attention mechanisms. These network structures could automatically represent n-grams and long-term dependency information, which helped the deep learning models gain better performance than those of conventional machine learning models (ie, logistic regression, naïve-Bayes, and support-vector machine) trained using the top of max-pooling embeddings [33,35]. Because of the complexity of the deep learning models, we were not able to precisely identify the decision mechanisms of the text semantics. However, according to the heatmaps of case-level word contribution, the potentially essential keywords helped identify SSIs in the form of distributed representation without any other handcrafted feature engineering or manual feature extraction. Potentially essential keywords included those suggesting endoscopic surgery, such as “laparoscope,” and “gallstone”; those suggesting clean surgery, such as “thyroid” and “centrum”; those suggesting colon surgery, such as “colon” and “transverse colon”; and those suggesting complex operation and prolonged operation time, such as “malignant,” “tuberculosis,” and “malignant tumor.”

Third, we tried many algorithmic techniques to avoid overfitting. For example, we used the L2 penalty, dropout, and early-stopping techniques. These techniques ensured the generalization ability of the model on different patient data to a certain extent.

Although we verified the effectiveness of the deep learning AMRAMS through both internal and external verification, many limitations still exist. The first limitation came from the training labels. The follow-up period of SSI by the infection prevention and control department was limited only to the hospitalization, which meant some of the SSIs that occurred after discharge might have been ignored. Although surgeons would conduct a careful examination of the incision before patient discharge, the occurrence of SSIs outside secondary care could not be completely eliminated. This bias would cause our model to underestimate the risk of patients developing SSIs.

The second limitation came from the patient population. Our dataset came from one medical site, and the time span of data collection was about 5 years, in which changes in patient population distribution, surgical procedures, and SSI prevention education and measures were inevitable. In our study, the internal validation results of the models were not completely consistent with the external verification results, which implied this point. If our model was to be applied to clinical practice, regular validation and update would be necessary.

The third limitation came from the missing values. We observed that many variables in our dataset had high missing rates, and many missing-value indicators contributed greatly to the model. In general, the missing data in the EMR system were not missing at random and were caused mainly by two reasons: inability to perform the measurement or a lack of indication to perform the measurement. For example, missing body weight information might indicate that the patient was unable to stand upright (eg, paralyzed) in order to measure the weight, whereas missing blood tests and liver function tests might suggest that the patient was healthy and young. In our study, we were not able to evaluate the potential influence of the missing data, because speculating the reason behind each missing value was complex and trivial. From a perspective of research, we could try to model the probability of missingness using other observed variables and conduct sensitivity analyses, which stimulate various missing patterns based on the predicted probability distributions, to evaluate the influences of missing data in our future studies. However, the ideal solutions to missing-data problems are still improving data quality and integrity in EMRs or developing less-biased imputation methods based on the patterns of the missing values in the patient records, the conditions of the patients, and the behaviors of the physicians.

The fourth limitation came from the models. We did not observe the attention mechanisms on the top of Bi-LSTM providing a great benefit over the convolutional kernels as claimed by Lin et al in their paper [35], probably because of the limited sizes of the training samples relative to the model parameters. Thus, we considered both self-attention and CNN as the best solutions in our current task. Because of the limitations of computing resources (ie, GPU [graphics processing unit] instances), we did not apply other state-of-the-art language models, such as BERT (bidirectional encoder representations from transformers) and its derivatives [41,42], to encode text information.

The fifth limitation came from the feature analysis. In this study, we only explored the correlations between the selected features and the occurrences of SSIs via the LASSO models, with detailed epidemic investigations and causal inferences remaining beyond the scope of this study. Moreover, because the LASSO models were trained using incomplete data and were not adjusted for potential confounders, the statistical inferences would be biased and the results would be hard to explain.

Future studies will focus on four points. First, we will try various new language models with deep transformers; encode various text information types, such as admission records, progress notes, and surgical records; and evaluate the models' performance. Second, we will confirm the effectiveness of our AMRAMS among multiple medical sites. Third, we will embed the AMRAMS into the EMR system and evaluate whether it can ultimately help reduce the occurrence of SSIs and optimize medical decision making. Fourth, our multimodal RAM solution could be validated for many other similar scenarios, such as syndromic or notifiable disease surveillance, adverse event monitoring, or ICD-10 coding support, in which both structured and free-text features would contribute to the judgement of final outcomes.

Conclusions

Our artificial intelligence–based multimodal risk assessment models for SSI based on EMR data and deep learning methods had significant advantages in terms of accuracy, compared with other conventional machine learning methods and the NNIS risk index. The semantic embeddings of clinical notes, whether generated using a simple max-pooling method or a deep learning method, improved the model performance further without any handcrafted feature engineering. Our models could replace the NNIS risk index to provide personalized guidance for the preoperative intervention of SSI. Through this case, we offered an easy-to-implement solution for building multimodal RAMs for similar scenarios, based on both structured and free-text data. Future studies should validate the generalization, reproducibility, and clinical impact in other medical settings.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The data description of the raw data and additional tables.
[DOCX File , 31 KB-Multimedia Appendix 1]

Multimedia Appendix 2

A portion of the patient records raw data.
[XLSX File (Microsoft Excel File), 4627 KB-Multimedia Appendix 2]

References

1. Cassini A, Plachouras D, Eckmanns T, Abu Sin M, Blank H, Ducomble T, et al. Burden of six healthcare-associated infections on European population health: Estimating incidence-based disability-adjusted life years through a population prevalence-based modelling study. PLoS Med 2016 Oct;13(10):e1002150 [FREE Full text] [doi: 10.1371/journal.pmed.1002150] [Medline: 27755545]
2. Lake JG, Weiner LM, Milstone AM, Saiman L, Magill SS, See I. Pathogen distribution and antimicrobial resistance among pediatric healthcare-associated infections reported to the National Healthcare Safety Network, 2011-2014. Infect Control Hosp Epidemiol 2018 Jan;39(1):1-11 [FREE Full text] [doi: 10.1017/ice.2017.236] [Medline: 29249216]
3. Hansen S, Schwab F, Zingg W, Gastmeier P, The Prohibit Study Group. Process and outcome indicators for infection control and prevention in European acute care hospitals in 2011 to 2012 - Results of the PROHIBIT study. Euro Surveill 2018 May;23(21):1-10 [FREE Full text] [doi: 10.2807/1560-7917.ES.2018.23.21.1700513] [Medline: 29845929]
4. Leaper D, Ousey K. Evidence update on prevention of surgical site infection. Curr Opin Infect Dis 2015 May;28(2):158-163. [doi: 10.1097/QCO.0000000000000144] [Medline: 25692267]
5. Waltz PK, Zuckerbraun BS. Surgical site infections and associated operative characteristics. Surg Infect (Larchmt) 2017;18(4):447-450. [doi: 10.1089/sur.2017.062] [Medline: 28448197]
6. Cossin S, Malavaud S, Jarno P, Giard M, L'Hériteau F, Simon L, ISO-RAISIN Steering Committee. Surgical site infection after valvular or coronary artery bypass surgery: 2008-2011 French SSI national ISO-RAISIN surveillance. J Hosp Infect 2015 Dec;91(3):225-230. [doi: 10.1016/j.jhin.2015.07.001] [Medline: 26321674]
7. Pollard TC, Newman JE, Barlow NJ, Price JD, Willett KM. Deep wound infection after proximal femoral fracture: Consequences and costs. J Hosp Infect 2006 Jul;63(2):133-139. [doi: 10.1016/j.jhin.2006.01.015] [Medline: 16621145]
8. Leaper D, Nazir J, Roberts C, Searle R. Economic and clinical contributions of an antimicrobial barrier dressing: A strategy for the reduction of surgical site infections. J Med Econ 2010;13(3):447-452. [doi: 10.3111/13696998.2010.502077] [Medline: 20653399]
9. Gheorghe A, Moran G, Duffy H, Roberts T, Pinkney T, Calvert M. Health utility values associated with surgical site infection: A systematic review. Value Health 2015 Dec;18(8):1126-1137 [FREE Full text] [doi: 10.1016/j.jval.2015.08.004] [Medline: 26686800]
10. Wang Z, Chen J, Wang P, Jie Z, Jin W, Wang G, et al. Surgical site infection after gastrointestinal surgery in China: A multicenter prospective study. J Surg Res 2019 Aug;240:206-218. [doi: 10.1016/j.jss.2019.03.017] [Medline: 30986636]
11. Zhou J, Ma X. Cost-benefit analysis of craniocerebral surgical site infection control in tertiary hospitals in China. J Infect Dev Ctries 2015 Mar 19;9(2):182-189 [FREE Full text] [doi: 10.3855/jidc.4482] [Medline: 25699493]

12. Fan Y, Wei Z, Wang W, Tan L, Jiang H, Tian L, et al. The incidence and distribution of surgical site infection in mainland China: A meta-analysis of 84 prospective observational studies. Sci Rep 2014 Oct 30;4:6783 [FREE Full text] [doi: 10.1038/srep06783] [Medline: 25356832]
13. Xiao Y, Shi G, Zhang J, Cao J, Liu L, Chen T, et al. Surgical site infection after laparoscopic and open appendectomy: A multicenter large consecutive cohort study. Surg Endosc 2015 Jul;29(6):1384-1393. [doi: 10.1007/s00464-014-3809-y] [Medline: 25303904]
14. Allegranzi B, Zayed B, Bischoff P, Kubilay NZ, de Jonge S, de Vries F, WHO Guidelines Development Group. New WHO recommendations on intraoperative and postoperative measures for surgical site infection prevention: An evidence-based global perspective. Lancet Infect Dis 2016 Dec;16(12):e288-e303. [doi: 10.1016/S1473-3099(16)30402-9] [Medline: 27816414]
15. Culver DH, Horan TC, Gaynes RP, Martone WJ, Jarvis WR, Emori TG, et al. Surgical wound infection rates by wound class, operative procedure, and patient risk index. National Nosocomial Infections Surveillance System. Am J Med 1991 Oct 16;91(3B):152S-157S. [doi: 10.1016/0002-9343(91)90361-z] [Medline: 1656747]
16. Mu Y, Edwards JR, Horan TC, Berrios-Torres SI, Fridkin SK. Improving risk-adjusted measures of surgical site infection for the National Healthcare Safety Network. Infect Control Hosp Epidemiol 2011 Oct;32(10):970-986. [doi: 10.1086/662016] [Medline: 21931247]
17. Grant R, Aupee M, Buchs NC, Cooper K, Eisenring M, Lamagni T, et al. Performance of surgical site infection risk prediction models in colorectal surgery: External validity assessment from three European national surveillance networks. Infect Control Hosp Epidemiol 2019 Sep;40(9):983-990. [doi: 10.1017/ice.2019.163] [Medline: 31218977]
18. van Walraven C, Musselman R. The Surgical Site Infection Risk Score (SSIRS): A model to predict the risk of surgical site infections. PLoS One 2013;8(6):e67167 [FREE Full text] [doi: 10.1371/journal.pone.0067167] [Medline: 23826224]
19. Bucher BT, Ferraro JP, Finlayson SR, Chapman WW, Gundlapalli AV. Use of computerized provider order entry events for postoperative complication surveillance. JAMA Surg 2019 Apr 01;154(4):311-318 [FREE Full text] [doi: 10.1001/jamasurg.2018.4874] [Medline: 30586132]
20. Pindyck T, Gupta K, Strymish J, Itani KM, Carter ME, Suo Y, et al. Validation of an electronic tool for flagging surgical site infections based on clinical practice patterns for triaging surveillance: Operational successes and barriers. Am J Infect Control 2018 Feb;46(2):186-190. [doi: 10.1016/j.ajic.2017.08.026] [Medline: 29031434]
21. Grundmeier RW, Xiao R, Ross RK, Ramos MJ, Karavite DJ, Michel JJ, et al. Identifying surgical site infections in electronic health data using predictive models. J Am Med Inform Assoc 2018 Sep 01;25(9):1160-1166. [doi: 10.1093/jamia/ocy075] [Medline: 29982511]
22. Colborn KL, Bronsert M, Amioka E, Hammermeister K, Henderson WG, Meguid R. Identification of surgical site infections using electronic health record data. Am J Infect Control 2018 Nov;46(11):1230-1235. [doi: 10.1016/j.ajic.2018.05.011] [Medline: 29907448]
23. The Ministry of Health of the People's Republic of China, No. 48 Decree. URL: http://www.gov.cn/ziliao/flfg/2006-07/25/content_344886.htm [accessed 2020-05-19]
24. Welch C, Petersen I, Walters K, Morris RW, Nazareth I, Kalaitzaki E, et al. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. Pharmacoepidemiol Drug Saf 2012 Jul;21(7):725-732. [doi: 10.1002/pds.2270] [Medline: 22052713]
25. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15;142(12):1255-1264. [doi: 10.1093/oxfordjournals.aje.a117592] [Medline: 7503045]
26. Santos I, Nedjah N, de Macedo Mourelle L. Sentiment analysis using convolutional neural network with fastText embeddings. In: Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). 2017 Nov 08 Presented at: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI); November 8-10, 2017; Arequipa, Peru p. 1-5. [doi: 10.1109/la-cci.2017.8285683]
27. Loshchilov I, Hutter F. arXiv. 2017 Nov. Decoupled weight decay regularization URL: https://ui.adsabs.harvard.edu/abs/2017arXiv171105101L [accessed 2020-05-24]
28. linguatools. Wikipedia monolingual corpora URL: https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/ [accessed 2020-05-19]
29. A-hospital. URL: http://www.a-hospital.com/ [accessed 2020-05-19]
30. GitHub. Jieba: Chinese text segmentation URL: https://github.com/fxsjy/jieba [accessed 2020-05-19]
31. Bojanowski P, Grave E, Joulin A, Mikolov T. arXiv. 2016 Jul. Enriching word vectors with subword information URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160704606B [accessed 2020-05-24]
32. Chen T, He T. Higgs boson discovery with boosted trees. In: Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS): 2014 Workshop in High Energy Physics and Machine Learning (HEPML). 2015 Presented at: 28th Conference on Neural Information Processing Systems (NIPS): 2014 Workshop in High Energy Physics and Machine Learning (HEPML); December 8-13, 2014; Montreal, Canada p. 69-80 URL: http://www.jmlr.org/proceedings/papers/v42/chen14.pdf
33. Kim Y. arXiv. 2014 Aug. Convolutional neural networks for sentence classification URL: https://ui.adsabs.harvard.edu/abs/2014arXiv1408.5882K [accessed 2020-05-24]

34. Hendrycks D, Gimpel K. arXiv. 2016 Jun. Gaussian Error Linear Units (GELUs) URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160608415H [accessed 2020-05-24]
35. Lin Z, Feng M, Nogueira dos Santos C, Yu M, Xiang B, Zhou B. arXiv. 2017 Mar. A structured self-attentive sentence embedding URL: https://ui.adsabs.harvard.edu/abs/2017arXiv170303130L [accessed 2020-05-24]
36. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 2005 Jan;16(1):73-81. [doi: 10.1097/01.ede.0000147512.81966.ba] [Medline: 15613948]
37. Tulio Ribeiro M, Singh S, Guestrin C. arXiv. 2016 Feb. Why should I trust you? Explaining the predictions of any classifier URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160204938T [accessed 2020-05-24]
38. Gong S, Guo H, Zhou H, Chen L, Yu Y. Morbidity and risk factors for surgical site infection following cesarean section in Guangdong Province, China. J Obstet Gynaecol Res 2012 Mar;38(3):509-515. [doi: 10.1111/j.1447-0756.2011.01746.x] [Medline: 22353388]
39. Gomila A, Carratalà J, Biondo S, Badia JM, Fraccalvieri D, Shaw E, VINCat Colon Surgery Group. Predictive factors for early- and late-onset surgical site infections in patients undergoing elective colorectal surgery. A multicentre, prospective, cohort study. J Hosp Infect 2018 May;99(1):24-30. [doi: 10.1016/j.jhin.2017.12.017] [Medline: 29288776]
40. Joulin A, Grave E, Bojanowski P, Mikolov T. arXiv. 2016 Jul. Bag of tricks for efficient text classification URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160701759J [accessed 2020-05-24]
41. Devlin J, Chang MW, Lee K, Toutanova K. arXiv. 2018 Oct. BERT: Pre-training of deep bidirectional transformers for language understanding URL: https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D [accessed 2020-05-24]
42. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. arXiv. 2019 Sep. ALBERT: A lite BERT for self-supervised learning of language representations URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190911942L [accessed 2020-05-24]

Abbreviations

AMRAMS: Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection
ASA: American Society of Anesthesiologists
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
Bi-LSTM: bidirectional long short-term memory
CNN: convolutional neural network
EMR: electronic medical record
GBDT: gradient boosting decision tree
GELU: Gaussian Error Linear Unit
GPU: graphics processing unit
HAI: health care–associated infection
ICD-10: International Classification of Diseases, Tenth Revision
LASSO: least absolute shrinkage and selection operator
LOS: length of stay
LSTM: long short-term memory
NHSN: National Healthcare Safety Network
NNIS: National Nosocomial Infections Surveillance
RAM: risk assessment model
ReLU: Rectified Linear Unit
ROC: receiver operating characteristic
SQL: Structured Query Language
SSI: surgical site infection
TF-IDF: term frequency–inverse document frequency

Edited by G Eysenbach; submitted 11.02.20; peer-reviewed by H Zhang, A Ćirković; comments to author 28.03.20; revised version received 15.04.20; accepted 19.04.20; published 15.06.20

Please cite as:
Chen W, Lu Z, You L, Zhou L, Xu J, Chen K
Artificial Intelligence–Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study
JMIR Med Inform 2020;8(6):e18186
URL: http://medinform.jmir.org/2020/6/e18186/
doi: 10.2196/18186
PMID: 32538798