Original Paper

Identifying the Risk of Sepsis in Patients With Cancer Using Digital Health Care Records: Machine Learning–Based Approach

Donghun Yang[1,2*], MSc; Jimin Kim[3*], PhD; Junsang Yoo[4], PhD; Won Chul Cha[4,5], MD, PhD; Hyojung Paik[2,3], PhD

[1]AI Technology Research Center, Division of S&T Digital Convergence, Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

[2]Department of Data and High Performance Computing Science, University of Science and Technology, Daejeon, Republic of Korea

[3]Center for Supercomputing Applications, Division of National Supercomputing, Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

[4]Department of Digital Health, Samsung Advanced Institute for Health Science & Technology, Sungkyunkwan University, Seoul, Republic of Korea

[5]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyojung Paik, PhD
Center for Supercomputing Applications
Division of National Supercomputing
Korea Institute of Science and Technology Information
245 Daehak-ro
Yuseong-Gu
Daejeon, 34141
Republic of Korea
Phone: 82 428690791
Email: hyojungpaik@gmail.com

Abstract

Background: Sepsis is diagnosed in millions of people every year, resulting in a high mortality rate. Although patients with sepsis present multimorbid conditions, including cancer, sepsis predictions have mainly focused on patients with severe injuries.

Objective: In this paper, we present a machine learning–based approach to identify the risk of sepsis in patients with cancer using electronic health records (EHRs).

Methods: We utilized deidentified anonymized EHRs of 8580 patients with cancer from the Samsung Medical Center in Korea in a longitudinal manner between 2014 and 2019. To build a prediction model based on physical status that would differ between sepsis and nonsepsis patients, we analyzed 2462 laboratory test results and 2266 medication prescriptions using graph network and statistical analyses. The medication relationships and lab test results from each analysis were used as additional learning features to train our predictive model.

Results: Patients with sepsis showed differential medication trajectories and physical status. For example, in the network-based analysis, narcotic analgesics were prescribed more often in the sepsis group, along with other drugs. Likewise, 35 types of lab tests, including albumin, globulin, and prothrombin time, showed significantly different distributions between sepsis and nonsepsis patients ($P<.001$). Our model outperformed the model trained using only common EHRs, showing an improved accuracy, area under the receiver operating characteristic (AUROC), and F1 score by 11.9%, 11.3%, and 13.6%, respectively. For the random forest–based model, the accuracy, AUROC, and F1 score were 0.692, 0.753, and 0.602, respectively.

Conclusions: We showed that lab tests and medication relationships can be used as efficient features for predicting sepsis in patients with cancer. Consequently, identifying the risk of sepsis in patients with cancer using EHRs and machine learning is feasible.

(JMIR Med Inform 2022;10(6):e37689) doi: 10.2196/37689

KEYWORDS

sepsis; cancer; EHR; machine learning; deep learning; mortality rate; learning model; electronic health record; network based analysis; sepsis risk; risk model; prediction model

XSL•FO
RenderX

Introduction

Sepsis is a life-threatening organ dysfunction in which a pathogen infection leads to a dysregulated host response to the infection [1]. Sepsis is diagnosed in millions of people every year globally, accounting for a high ratio of in-hospital mortality (25%-50%) [2]. In particular, the mortality rate increases dramatically when septic shock is established [3,4]. Although a timely diagnosis of sepsis is essential for a promising prognosis, only minor cold-like symptoms, such as fever, excessive breathing, and increased pulse rate, are presented in the early stage of sepsis [5]. Therefore, in hospitals, patients admitted to the ward may suffer from septic shock after clinicians have missed the signature symptoms of sepsis. Thus, it is important to stratify high-risk patients and provide appropriate treatment in a short amount of time [6].

Sepsis has shown a substantial incidence in patients with low immunity, such as patients with cancer, patients who are elderly, and newborns [7]. Patients with cancer are at high risk for sepsis, as many are immunosuppressed due to the cancer itself and chemotherapy treatment [8]. For example, leukocyte counts are lowered, especially when anticancer treatments decrease bone marrow function, suppressing immune response to the pathogen [9]. Although predicting sepsis in patients with cancer is essential, an early identification of the risk of sepsis remains an unmet medical need.

Various studies have been conducted to identify the risk of sepsis, including a statistical model–based approach for emergency room (ER) patients [10], a machine learning–based approach for inpatients [11], and an approach using unstructured clinical data [12]. The majority of previous studies have focused on patients with severe trauma in the intensive care unit (ICU). However, the stratification of sepsis risk among patients with cancer has scarcely been conducted.

Our study aimed to predict the risk of sepsis in patients with cancer at an early stage using clinical information and a machine learning approach. We utilized the deidentified electronic health records (EHRs) from the Samsung Medical Center (SMC) in Korea of 8580 patients with cancer, including inpatients, outpatients, ICU patients, and ER patients. Drug prescriptions and laboratory test results are known to reflect the physical status of patients [13]. In our previous study, we showed that distributions of lab test results recapitulate the physical states of patients, including disease signatures and drug-associated responses [14]. Prescriptions of medications for cancer are mainly determined based on the patients' medical conditions. Thus, we hypothesized that the patterns of prescribed medications and lab test results would be different between the sepsis and nonsepsis groups. To validate our hypothesis, we analyzed 2462 lab test results and 2266 medication prescriptions using network-based association rule [15] analysis and statistical analysis.

Based on the results of the analyses, we propose a machine learning–based sepsis predictive model that can reflect the physical conditions of patients with cancer and is trained on the prescribed drug and lab test patterns as well as EHRs, which are widely used in the reported sepsis prediction approaches [16,17].

Methods

Study Sample

Data were prepared from the Clinical Data Warehouse (CDW) and the SMC cancer registry, Seoul, South Korea, and deidentification was performed on the collected data. The study population included adult patients diagnosed with lung, liver, and breast cancer who visited the ER within 5 years of being diagnosed with cancer. The inclusion criteria were patients with cancer registered at the study sites. Patients were excluded from the study cohort if they met the following exclusion criteria: those under 18 years of age, those with multiple cancers, those who had not visited the emergency room within 5 years after the first cancer diagnosis, and those with ICD-10 codes not matched with C22, C34, and C50. The data were constructed by reflecting various EHR information such as hospitalization data, diagnosis code of cancer or other underlying disease, vital signs, genomic information, medication prescription, surgical history, radiation treatment, and lab test information for 5 years (2014-2019) before and after the cancer diagnosis of 8580 patients with cancer, including inpatients, outpatients, ICU patients, and ER patients. Most of the currently published sepsis prediction models use information within 48 hours before the onset of sepsis. However, due to the high risk of sepsis, it was considered necessary to predict in advance, so information 2 days prior to the ER visit was used. Data earlier than 7 days were somewhat difficult to consider as having an effect on the onset of sepsis, so the filtration criteria was set to 2-7 days.

Ethics Approval

The institutional ethics committee of SMC approved this study (Institutional Review Board File #2019-06-071).

Identifying Patients With Sepsis

We identified patients with sepsis using the Sequential Organ Failure Assessment (SOFA) scores of Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) guidelines [18] for a total of 18,610 ER visits by 8580 patients with cancer using the following procedures:

1. Nursing records, inspection records, clinical information, and medication prescription data were extracted from the CDW.
2. The variables were preprocessed to obtain the SOFA scores.
3. SOFA scores for each patient were calculated each time.
4. The time window was set by checking whether antibiotics were administered intravenously within 24 hours before and after the bacterial culture test.
5. Patients with sepsis were identified if their SOFA score changed by 2 points or more within the time window.
6. In accordance with the Sepsis-3 guidelines, if the SOFA score could not be measured in advance, it was considered 0. Consequently, if the change in the SOFA score was 2 or higher in the first visit to the ER, the patient was considered to be experiencing sepsis.

XSL•FO
RenderX

Data Filtering and Preprocessing

We aligned the collected EHRs of 8580 patients with cancer based on the date of the ER visit and filtered patients with information 2-7 days prior to the ER visit. Each patient's diagnostic code was recorded as an ICD-9 code and standardized to 3 digits for use as a categorical feature. Because there was a possibility of information leakage from giving hints to the machine learning–based predictive model, lab test results centered on specific disease groups were removed, and only the lab test information performed on over 60% of the patients was used. All categorical features were preprocessed using one-hot encoding, and all binary categorical features were encoded as 0 or 1. In addition, missing values were imputed with the mean value of patients with the same type of cancer, the same sex, and the same age, and extreme outlier values were removed.

Graph Network–Based Association Rule Analysis

Graph network–based association rules were performed on 2266 drug prescriptions. An association rule is a method for discovering frequent patterns and relationships between items from complicated data and can be employed to conceptualize complex dynamic systems comprising each interacting event [15]. Using the frequent pattern growth (FP-growth) algorithm [19], frequent relationships of drugs prescribed on the same day were analyzed. Next, only the group sets with a minimum support value of 0.05 or greater were selected. The support value ($S(D_i \rightarrow D_j)$), defined as in Equation (1), implied how often the sets go together when items are being tied up simultaneously, where *N(s)* represents the total number of prescriptions, and *N* (D_i,D_j) represents the number of events in which the *i-th* and *j-th* drugs were prescribed on the same day.

$$S(D_i \rightarrow D_j) = \frac{N(D_i, D_j)}{N(s)}$$

Finally, after designating each selected drug as a node, we plotted a graph network to visualize the result of the association rule analysis. The edges depicted the correlations of each drug.

Vectorization of Prescribed Medication Relationships

We vectorized the relationships found through graph network–based association rule analysis to be used as an input for the machine learning–based sepsis prediction model. After multiplying each one-hot encoded drug selected through the aforementioned analysis by the number of prescription days, the relationship for each pair of values was vectorized using the 3 formulas proposed in our previous study [20]. These 3 formulas ($r(I, H, T)$) comprised the interaction (*I*), the harmonized average (*H*), and the arctangent (*T*), in which (*I*) determined the level of interaction, (*H*) determined the overall intensity in a sensitive manner, and (*T*) determined the geometric angle difference as a single scalar value for each pair, defined as in Equation (2), where $D_i^{(p,s)}$ and $D_j^{(p,s)}$ indicate the *i-th* and *j-th* drug of the *s-th* prescriptions for the *p-th* patient, respectively. The *D* value represents the prescription frequency of each medication.

$$r\left(I_{ij}^{(p,s)}, H_{ij}^{(p,s)}, T_{ij}^{(p,s)}\right) = \begin{cases} I_{ij}^{(p,s)} = D_i^{(p,s)} \cdot D_j^{(p,s)} \\ H_{ij}^{(p,s)} = \left(\frac{2}{(D_i^{(p,s)})^{-1} + (D_j^{(p,s)})^{-1}}\right) \\ T_{ij}^{(p,s)} = \tan^{-1}(D_i^{(p,s)}/D_j^{(p,s)}) \cdot (180°/\pi) \end{cases}$$

Prediction of Sepsis Using Machine Learning Approaches

We trained models on vectorized drug relationships and selected lab test types, along with the common EHRs that are widely used in the reported sepsis prediction models [16,17]. We considered 2 machine learning models comprising logistic regression (LR) [21] and random forest [22] and 3 deep learning models comprising artificial neural networks (ANNs) [23], residual convolutional neural networks (ResNet10) [24], and long short-term memory recurrent neural networks (RNN-LSTMs) [25]. When applied to the model, the data were reshaped to (1, 42, 42) for the ResNet10 model and padded to the maximum length of the sequence and reshaped to (number of patients, time sequence, number of features) for the LSTM model. We investigated the important features using Shapley Additive Explanations (SHAP) [26]. SHAP, one of the Explainable Artificial Intelligence (XAI) techniques, is a method used to interpret results from deep learning and machine learning models and is based on game theory. We used Tree SHAP explainer to calculate the Shapley values.

All proposed approaches were implemented using the Python 3.7 library, such as PyTorch 1.5, Scikit-learn, and SHAP, on an NVIDIA TITAN RTX 24 GB × 2. The source code is available on GitHub [27].

Results

Characteristics of the Filtered and Preprocessed Data Set From SMC

The overall process of our study is shown in Figure 1. We analyzed data from 8580 patients obtained from the CDW of SMC. Using the SOFA scores of the Sepsis-3 guidelines, of a total of 18,610 ER visits by 8580 patients with cancer, 2960 visits were identified as sepsis and 15,650 visits as nonsepsis. As a result of filtering the patients, the control group included 928 patients, and the sepsis group included 455 patients. The statistics of the filtered and preprocessed data set that was used to build the sepsis predictive model are shown in Table 1.

In the control group (ie, nonsepsis patients with cancer), there were 490 (52.8%) males and 438 (47.2%) females. The mean age was 58.2 (SD 11.0) years, and the average weight was 63.7 (SD 10.7) kg. In terms of the initial cancer diagnosis of each patient, 180 (19.4%) had liver cancer, 533 (57.4%) had lung cancer, and 215 (23.2%) had breast cancer. Meanwhile, in the sepsis group, there were 324 (71.2%) males and 131 (28.8%) females, with a relatively higher proportion of males than the control group. The mean age of the sepsis group was 60.3 (SD 0.5) years, and the average weight was 64.3 (SD 11.3) kg. In the sepsis group, 140 (30.8%) patients had liver cancer, 274 (60.2%) had lung cancer, and 41 (9%) had breast cancer. With these prepared data sets from SMC, we analyzed the differences in medication patterns by group.

XSL•FO
RenderX

Figure 1. Study overview. CDW: Clinical Data Warehouse; EHR: electronic health record; ER: emergency room; ER visits: total number of ER visits by the patients; SOFA: Sequential Organ Failure Assessment of the Sepsis-3 guidelines; SMC: Samsung Medical Center.

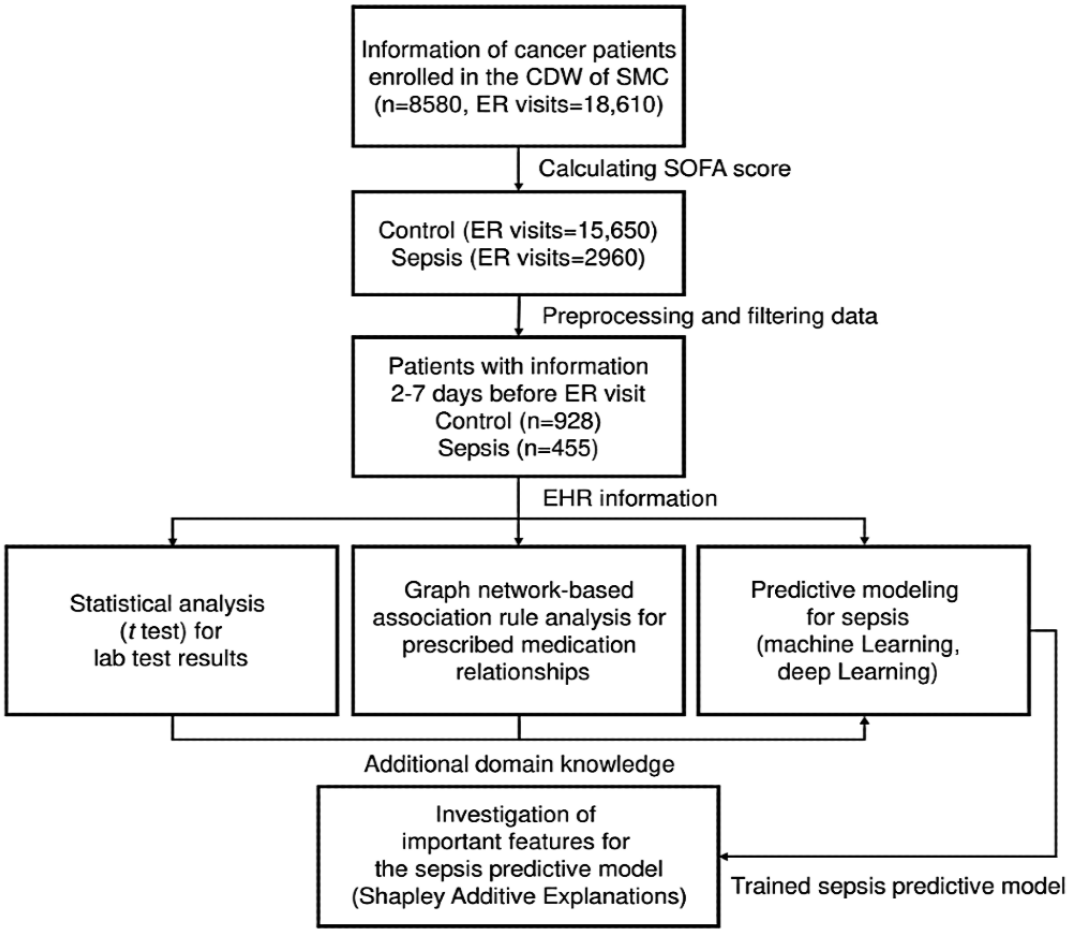


Table 1. Statistics of the input data used to build the sepsis predictive model.

Patient characteristics	Total (N=1383)	Control group (n=928)	Sepsis group (n=455)
Sex, n (%)			
Male	814 (58.9)	490 (52.8)	324 (71.2)
Female	569 (41.1)	438 (47.2)	131 (28.8)
Age (years), mean (SD)	58.9 (10.9)	58.2 (11)	60.3 (0.5)
Weight (kg), mean (SD)	63.9 (0.9)	63.7 (10.7)	64.3 (11.3)
Cancer, n (%)			
Liver	320 (23.1)	180 (19.4)	140 (30.8)
Lung	807 (58.4)	533 (57.4)	274 (60.2)
Breast	256 (18.5)	215 (23.2)	41 (9.0)
Emergency room visits, n (%)	1466 (100)	991 (68)	475 (32)

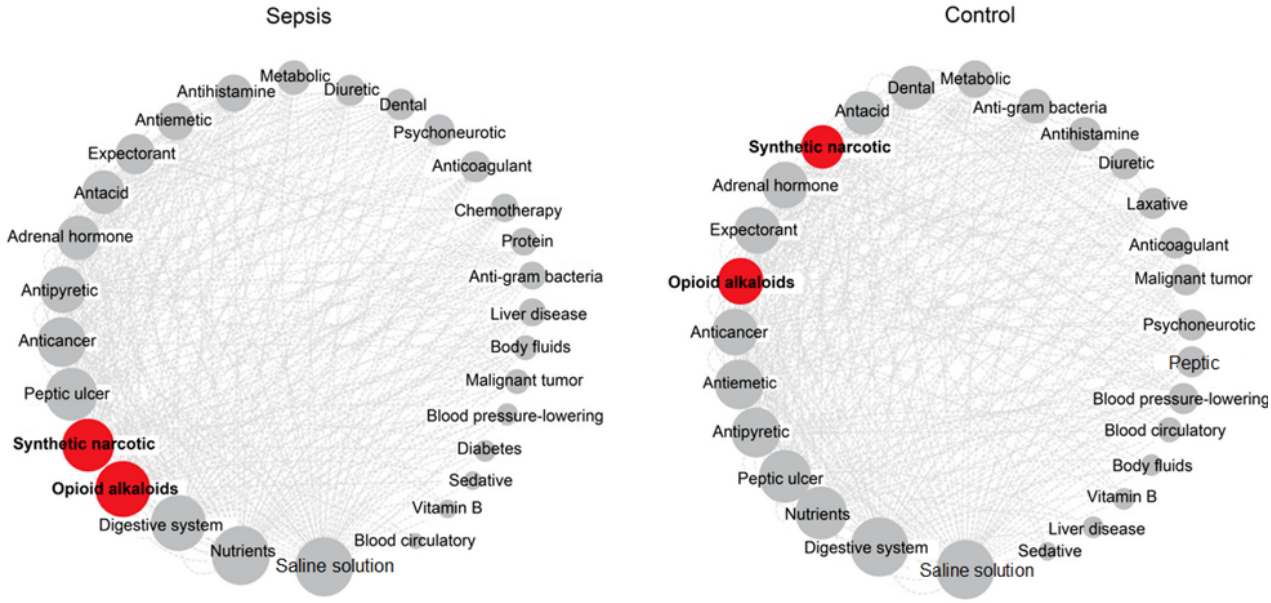
Graph Network–Based Association Analysis for Prescribed Medications

Using the FP-growth algorithm, we analyzed patterns of the medications prescribed on the same day in 2666 prescriptions from the preprocessed and filtered EHR data. According to the analysis results, only group sets with a minimum support value of 0.05 or greater were selected. Of a total of 101 different drug types, 406 relationships among 29 drugs and 378 relationships among 28 drugs were selected for the sepsis group and nonsepsis group, respectively. To visualize the associations between the drug prescriptions, we constructed 2 graph networks with nodes representing the selected drugs and edges depicting the relationships among the nodes (Figure 2). The size of a node was determined by its average shortest path distance (Multimedia Appendix 1, graph A) and the number of edges (Multimedia Appendix 1, graph B), representing the topological properties of the network. A larger node meant that the corresponding drug was prescribed more often with other drugs compared to small nodes.

XSL•FO
RenderX

The patterns of medications between the sepsis and control groups were different. Nodes for medications such as "saline solution," which are commonly prescribed for most patients, showed similar patterns in both networks (Figure 2), whereas "opioid alkaloids" and "synthetic narcotic" nodes were ranked higher in the sepsis group than in the control group (Multimedia Appendix 1, graphs A and B). These kinds of narcotic analgesic nodes were the bottleneck and central nodes, meaning that they were prescribed more often with other drugs in the sepsis group. Thus, we were able to confirm our hypothesis that relationships and patterns of prescribed medications were distinct in the 2 groups.

Figure 2. Results of the graph network-based association analysis for prescribed medications. Graph networks for sepsis (n=29) and control (n=28) patients. Each node represents the medication selected through association rule analysis (support value ≥ 0.05), and edges depict that the linked nodes were prescribed together. Red nodes and bars represent the drugs with different patterns between the sepsis and control patients.



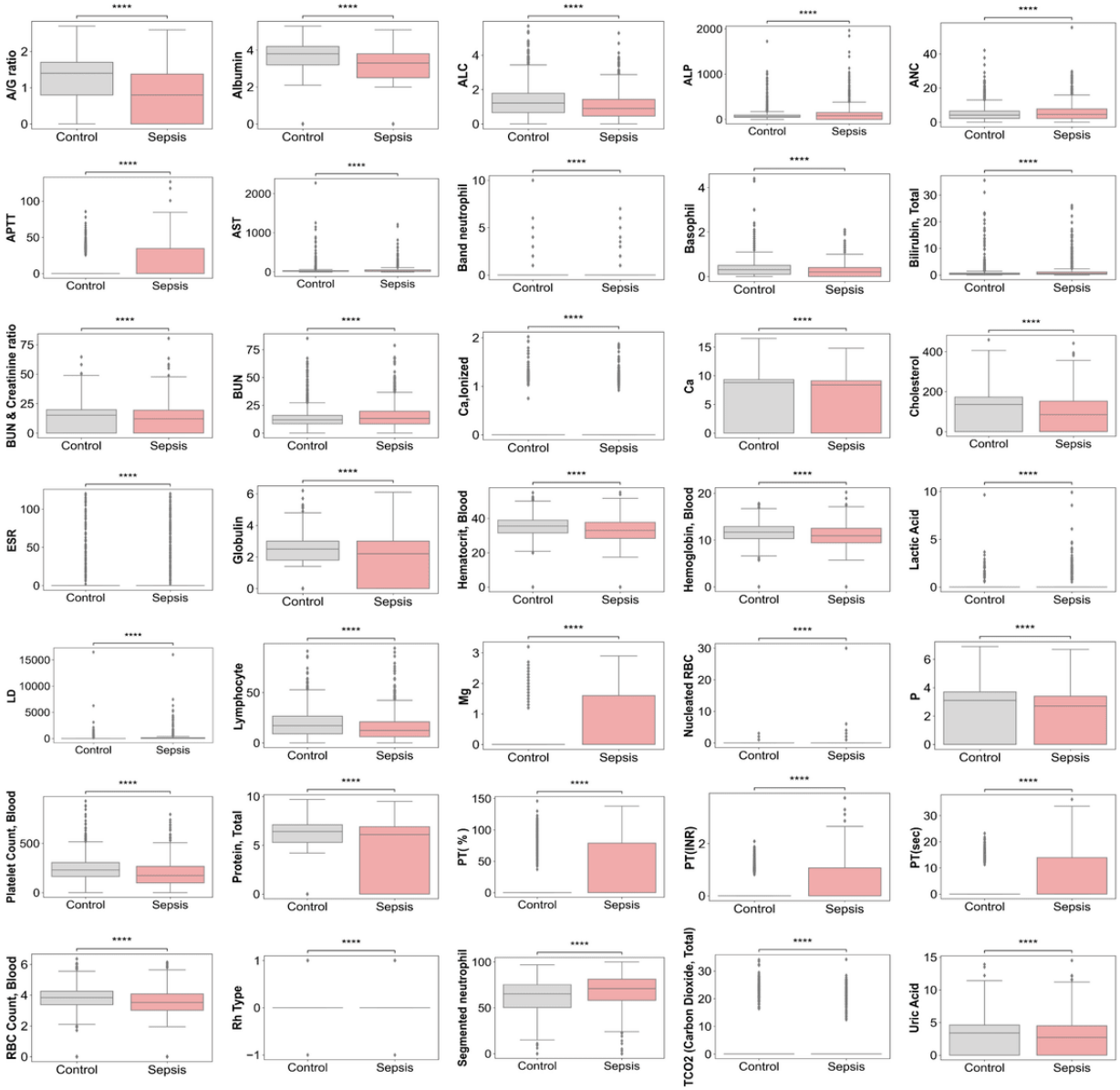
Statistical Analysis for Lab Tests

Using a *t* test, we analyzed the 2462 lab test results of the preprocessed and filtered EHR data to find lab test items with significantly different distributions between the sepsis and nonsepsis groups. Multimedia Appendix 2 presents the means and standard deviations of the 2 groups for all lab test types, where the *P* is symbolized (no significance: NS, *P*<.05: *, *P*<.001: ***, *P*<.005: **, *P*<.001: ****).

Figure 3 presents the distributions of the 2 groups for our selected lab test types, including predictors that are well-known hallmarks of sepsis. The changes in albumin, total protein, and cholesterol levels reflected the higher risk of mortality in patients with sepsis [28], and in our results, the values of these factors were significantly lower in patients with sepsis, with an albumin level of 2.73 (SD 1.57) versus 3.3 (SD 1.47), total protein of 4.77 (SD 2.98) versus 5.27 (SD 2.82), and a cholesterol level of 81.06 (SD 84.79) versus 111.46 (SD 84.3), respectively, with *P*<.001. A decreased albumin/globulin ratio (A/G ratio) was recently reported as a novel independent predictor of the development of postflexible ureteroscopic sepsis [29]. In this study, the A/G ratio in the sepsis group was lower than that in the nonsepsis group, at 0.76 (SD 0.7) versus 1.14 (SD 0.69), respectively, with *P*<.001. Furthermore, several studies found that activated partial thromboplastin time (APTT) and prothrombin time (PT) [30,31] are prognostic biomarkers for the specific identification of patients with sepsis. We observed that the occurrence of sepsis led to increased APTT and PT values, with an *APTT* value of 6.89 (SD 15.01) to 14.06 (SD 19.72), *PT(%)* value of 17.51 (SD 35.82) to 33.27 (SD 42.26), *PT(INR)* value of 0.23 (SD 0.46) to 0.49 (SD 0.63), and *PT(sec)* value of 2.9 (SD 5.84) to 6.14 (SD 7.69), with *P*<.001. With favorable consistency between the identified lab test differences of our candidates and known biomarker candidates of sepsis, we felt confident that the SMC EMRs successfully recapitulated the physical and biological signatures of the patients with sepsis. In other words, these results suggested that the 35 selected biomarkers could characteristically reflect the biological features of the patients with sepsis. Thus, we utilized these 35 lab test results as learning features to establish our prediction model for sepsis, as shown in Figure 3.

XSL•FO
RenderX

Figure 3. Distributions of the two groups for the selected lab test types. Of the 64 total lab test types (Multimedia Appendix 1), 35 lab test types showed significantly different distributions between the sepsis and control groups. ****: *P*<.001; A/G ratio: albumin/globulin ratio; ALC: absolute lymphocyte count; ALP: alkaline phosphatase; ANC: absolute neutrophil count; APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; ESR: erythrocyte sedimentation rate; LD: lactate dehydrogenase; PT: prothrombin time; RBC: red blood cell.



Prediction of Sepsis Using Machine Learning Approaches

Using vectorized drug relationships and the values of the selected lab test types along with common EHRs, we trained 2 machine learning models (logistic regression and random forest) and 3 deep learning models (ANN, convolutional neural network [CNN], and RNN) to build a sepsis prediction model based on the physical status of patients with cancer. A total of 465 relationships between 31 drugs selected through association rule analysis were vectorized using the 3 formulas described in the Methods section. A total of 1395 (465 × 3) drug relationship vectors, the values of the selected 35 lab test types, and common EHR information including anonymized personal information, hospitalization data, and cancer diagnosis code were used as inputs for model training. We used simple logistic regression (LR) and RNN-LSTM for the logistic regression–based and RNN-LSTM–based models, respectively. The random forest–based model comprised 20 trees, and the ANN-based model comprised input and output layers, as well as hidden layers. In addition, we used ResNet10 consisting of 10 convolution layers, fully connected layers, and residual connections for the CNN-based model. All hyperparameters, such as the number of trees in the random forest model, batch size, learning rate, and number of layers in the deep learning models, were selected as optimal values for each model through grid searches [32]. The list of feature variables used in our proposed model is given in Multimedia Appendix 3. To verify the proposed sepsis prediction model, we compared the predictive performances with the models trained on only

XSL•FO
RenderX

common EHRs (ie, demography, diagnoses codes, and others) and the models trained on common EHRs and drug relationships by 5-fold cross-validation. Regarding performance evaluation metrics, the accuracy, area under the receiver operating characteristic (AUROC), area under the precision-recall curve (AUPRC), precision, recall, and F1 score were used. Multimedia Appendix 4 shows the performance evaluation results.

The overall performance of the proposed models with EHRs, lab data, and drug relationships were superior to that of the other models. The proposed random forest–based model showed the highest value in all the evaluation metrics except for recall (accuracy: 0.692, AUROC: 0.753, AUPRC: 0.573, precision: 0.518, recall: 0.718, and F1 score: 0.602). In the case of recall, the proposed ANN-based model showed the highest value (accuracy: 0.654, AUROC: 0.723, AUPRC: 0.522, precision: 0.477, recall: 0.721, and F1 score: 0.574). In particular, the proposed random forest–based proposed model recorded the largest performance improvement in all the metrics compared to the model trained on drug relationships and common EHRs (accuracy: 0.645 to 0.692, AUROC: 0.69 to 0.753, AUPRC: 0.487 to 0.573, precision: 0.465 to 0.518, recall: 0.629 to 0.718, and F1 score: 0.534 to 0.602). In addition, the proposed RNN-LSTM–based model showed the greatest performance improvement in accuracy, the AUROC, the AUPRC, and precision (accuracy: 0.603 to 0.675, AUROC: 0.655 to 0.729, AUPRC: 0.447 to 0.555, and precision: 0.43 to 0.504), and the ResNet10-based model showed the highest improvement in recall and the F1 score (recall: 0.577 to 0.689, F1 score: 0.499 to 0.567) compared to the model trained on only common EHRs. These findings suggest that the drug relationships and the selected lab test types were the main contributors to the proposed sepsis predictive models for patients with cancer.

Investigation of Important Features

To evaluate the contributions of the learning features, SHAP, an XAI technique, was utilized for the proposed random forest–based model, which showed the best performance when investigating important features that contributed to the prediction. Multimedia Appendix 5 shows the contribution ratios of the top 50 important features among 1738 features obtained through SHAP, where the x-axis denotes the feature contribution ratio, and the y-axis denotes the names of the features.

The top 50 important features include 26 lab test types and 15 drug relationships among the 31 drugs and the 35 lab test types selected by *t* test and association rule analysis, respectively. The 15 drug relationships contained narcotic analgesic drugs such as "*opioid alkaloids*" and "*synthetic narcotics*," which were prescribed more, along with other drugs, in the sepsis group. Among the characteristics of the patients with cancer, the number of cancer-infiltrating lymph nodes (Ca_LN_no), the degree of cancer extent (Extend_CD), and the size of the primary tumor (T_CD) were observed as decisive contributing factors.

As expected, prognostic biomarkers of sepsis, such as the albumin level, PT, A/G ratio, total protein level, and cholesterol level, ranked high. The blood platelet count has also been identified as a major contributor, and platelets are involved in mechanisms that promote immune responses and coagulation activation. Thrombocytopenia is common in ICU patients with sepsis and is reportedly associated with fatal outcomes [33]. The migration of neutrophils to infection sites is essential in the host's defense against invading pathogens during sepsis [34], which may have led to the absolute neutrophil count or segmented neutrophils improving the predictive performance of the model. Moreover, when expanded to the top 100, all selected lab test types except "*band neutrophil*," "*nucleated RBC*," and "*carbon dioxide, total*," as well as 49 drug relationships comprising 22 selected drugs, were included. These results show that the selected drug relationships and lab tests were important features in the proposed sepsis predictive model, suggesting that these features contributed to the accurate prediction of the model.

Discussion

Principal Findings

This study presents a machine learning–based approach to identify sepsis risk in patients with cancer at an early stage (2 days before onset). We elucidated that the relationships of prescribed medications and lab test patterns were distinct in the sepsis and control groups. Based on these analysis results, we built a machine learning–based sepsis prediction model trained on lab test items and vectorized drug relationships, along with EHRs. The proposed model outperformed the model trained on medication relationships or common EHRs. In particular, the proposed random forest–based model showed the best sepsis prediction performance (accuracy: 0.692, AUROC: 0.753, and F1 score: 0.602) and showed the greatest performance improvement. Furthermore, we demonstrated that the selected lab test results and drug relationships were indeed important features and mainly contributed to the accurate prediction of our proposed model. Therefore, lab tests and medication relationships can be used as efficient features for predicting sepsis. Consequently, it will be possible to use EHR information and deep learning methods to identify the risk of sepsis in patients with cancer.

Limitations

Several limitations of the study should be noted. First, health records are not intended specifically for research; nonbilling-related data, including self-reported data such as smoking status, would be partially inaccurate. As depicted in Table 1, a substantial portion of patients with cancer are diagnosed with liver or lung cancer. Although there is a fairly significant incidence of liver and lung cancer in Korea [35], characteristic signatures of lab results and medications (eg, a lower A/G ratio and usage of opioid alkaloids) among patients with sepsis should be addressed at the pan-cancer level in further studies. For the contribution of the relationships of medication pairs, we acknowledge that there are many stakeholders in the prescription of medications, including insurance coverage. In this study, patients in Korea were all covered by the National Health Insurance Service. Thus, there would be a limited utilization of the relationship of medication combinations for model training in further applications from different countries corresponding to the heterogeneous milieu of insurance coverages.

XSL•FO
RenderX

As we hypothesized, our network-based analysis disclosed distinct patterns of medications used between sepsis and nonsepsis patients with cancer. For example, synthetic narcotics and opioid agents appeared to be more frequently prescribed with other agents. These features (ie, lab test results and medication patterns) mainly contributed to the high performance of our prediction model. Because the usage of opioids is a known risk factor for sepsis [36], the possibility of iatrogenic effects for the medication pattern–based prediction of sepsis in patients with cancer remains unclear. Therefore, drug-drug interactions between synthetic narcotics and anticancer agents should be addressed to further understand sepsis in patients with cancer. The retrospective analysis of EHRs paves the way for future research to understand sepsis among patients with cancer.

Conclusion

To our knowledge, previous prognostic evaluation tools and models primarily use patient information obtained after admission to the ICU, and there are many limitations for medical interventions. However, since most patients with cancer are hospitalized through the emergency room for the initial diagnosis of sepsis, an appropriate evaluation tool is needed to identify the risk in advance. This study can be referenced as a baseline for efficiently predicting the onset of sepsis in patients with cancer, and the model is expected to be able to identify sepsis risk more accurately and earlier than before in the medical field.

Acknowledgments

This work was supported by the Korea Institute of Science and Technology Information (KISTI) (K-21-L02-C10, K-20-L02-C10-S01). Authors HP and JK were also supported by the Ministry of Science and ICT (N-21-NM-CA08-S01). This research was also supported by the Program of the National Research Foundation (NRF) funded by the Korean government (2021M3H9A203052011). The computational analysis was supported by the National Supercomputing Center, including the resources and technology. We also thank Samsung Medical Center for providing the data.

Data Availability

The data sets used and analyzed in this study are available from the corresponding author on reasonable request.

Authors' Contributions

HP conceptualized the study and its methodology, supervised the study, and was responsible for funding acquisition. WCC and JY acquired the data. DY and JK carried out the formal analysis, developed and validated the model, and carried out the visualization. WCC and HP were responsible for the project administration. JY performed a technical review of the Methods section. DY, JK, and HP wrote the original draft of the paper, and JK and HP reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Topological properties of the network.
[PNG File , 233 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Comparison of lab test numerical distribution in the sepsis versus control groups.
[DOCX File , 23 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Description of the feature variables used in the proposed sepsis prediction model.
[DOCX File , 22 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Performance evaluation results. (A-F) Comparison of the sepsis prediction performance for the models trained on only common EHRs (only EHRs), the models trained on common EHRs and drug relationships (EHRs+Drug Rel), and the models trained on drug relationships and lab test types, along with common EHRs (EHRs+Drug Rel+Lab test types (proposed)) obtained by 5-fold cross-validation. AUROC: area under the receiver operating characteristic; AUPRC: area under the precision-recall curve; EHR: electronic health record; LR: logistic regression; RF: random forest; ANN: artificial neural network; ResNet10: residual convolutional neural network with 10 layers; LSTM: long short-term memory recurrent neural network.
[PNG File , 106 KB-Multimedia Appendix 4]

XSL•FO
RenderX

Multimedia Appendix 5

Feature contribution ratio of the Shapley Additive Explanations. Bar chart shows the feature contribution ratio (%) of the top 50 features obtained by the Shapley Additive Explanations algorithm for the random forest-based sepsis prediction model. A/G ratio: albumin/globulin ratio; ALP: alkaline phosphatase; AM: taking medications in the morning (ante meridiem); APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; Ca_LN_no: number of cancer-infiltrating lymph nodes; Extend_CD: degree of cancer extent; IV: intravenous administration; T_CD: size of primary tumor; PT: prothrombin time; RBC: red blood cell.
[PNG File , 132 KB-Multimedia Appendix 5]

References

1. Gullo A, Bianco N, Berlot G. Management of severe sepsis and septic shock: challenges and recommendations. Crit Care Clin 2006 Jul;22(3):489-501, ix. [doi: 10.1016/j.ccc.2006.03.006] [Medline: 16893735]
2. Fleischmann C, Scherag A, Adhikari N, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. Am J Respir Crit Care Med 2016 Feb 01;193(3):259-272. [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]
3. Martin GS. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert Rev Anti Infect Ther 2012 Jun 10;10(6):701-706 [FREE Full text] [doi: 10.1586/eri.12.50] [Medline: 22734959]
4. American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Crit Care Med 1992;20(6):864-874. [doi: 10.1097/00003246-199206000-00025]
5. Hunt A. Sepsis: an overview of the signs, symptoms, diagnosis, treatment and pathophysiology. Emerg Nurse 2019 Sep 02;27(5):32-41. [doi: 10.7748/en.2019.e1926] [Medline: 31475503]
6. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. Crit Care Med 2006;34(6):1589-1596. [doi: 10.1097/01.ccm.0000217961.75225.e9]
7. Iskander KN, Osuchowski MF, Stearns-Kurosawa DJ, Kurosawa S, Stepien D, Valentine C, et al. Sepsis: multiple abnormalities, heterogeneous responses, and evolving understanding. Physiol Rev 2013 Jul;93(3):1247-1288 [FREE Full text] [doi: 10.1152/physrev.00037.2012] [Medline: 23899564]
8. Ménétrier-Caux C, Ray-Coquard I, Blay J, Caux C. Lymphopenia in Cancer Patients and its Effects on Response to Immunotherapy: an opportunity for combination with Cytokines? J Immunother Cancer 2019 Mar 28;7(1):85 [FREE Full text] [doi: 10.1186/s40425-019-0549-5] [Medline: 30922400]
9. Williams MD, Braun L, Cooper LM, Johnston J, Weiss RV, Qualy RL, et al. Hospitalized cancer patients with severe sepsis: analysis of incidence, mortality, and associated costs of care. Crit Care 2004 Oct;8(5):R291-R298 [FREE Full text] [doi: 10.1186/cc2893] [Medline: 15469571]
10. Brown SM, Jones J, Kuttler KG, Keddington RK, Allen TL, Haug P. Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department. BMC Emerg Med 2016 Aug 22;16(1):31 [FREE Full text] [doi: 10.1186/s12873-016-0095-0] [Medline: 27549755]
11. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019 Jun;109:79-84 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.04.027] [Medline: 31035074]
12. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711 [FREE Full text] [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
13. Park MY, Yoon D, Lee K, Kang SY, Park I, Lee S, et al. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. Pharmacoepidemiol Drug Saf 2011 Jun 06;20(6):598-607. [doi: 10.1002/pds.2139] [Medline: 21472818]
14. Paik H, Chung A, Park H, Park RW, Suk K, Kim J, et al. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. Sci Rep 2015 Mar 05;5(1):8580 [FREE Full text] [doi: 10.1038/srep08580] [Medline: 25739475]
15. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Data Bases. 1994 Presented at: VLDB '94; Sept 12-15; Santiago de Chile, Chile.
16. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2020 Feb;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
17. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020 Mar 21;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]

XSL•FO
RenderX

18. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
19. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. SIGMOD Rec 2000 Jun;29(2):1-12. [doi: 10.1145/335191.335372]
20. Kim B, Kim Y, Park CHK, Rhee SJ, Kim YS, Leventhal BL, et al. Identifying the medical lethality of suicide attempts using network analysis and deep learning: nationwide study. JMIR Med Inform 2020 Jul 09;8(7):e14500 [FREE Full text] [doi: 10.2196/14500] [Medline: 32673253]
21. Cox DR. The Regression Analysis of Binary Sequences. J R Stat Soc Series B Stat Methodol 2018 Dec 05;21(1):238-238. [doi: 10.1111/j.2517-6161.1959.tb00334.x]
22. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
23. Tadeusiewicz R. Neural networks: A comprehensive foundation. Control Eng Pract 1995 May;3(5):746-747. [doi: 10.1016/0967-0661(95)90080-2]
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016 Jun Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
25. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of the Annual Conference of the International Speech Communication Association. 2014 Feb Presented at: INTERSPEECH; Sep 14-18; Singapore.
26. Lundberg S, Lee S. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017.
27. Deep Sepsis in Cancer. URL: https://github.com/yangdonghun3/deep_sepsis_in_cancer [accessed 2022-05-24]
28. Takegawa R, Kabata D, Shimizu K, Hisano S, Ogura H, Shintani A, et al. Serum albumin as a risk factor for death in patients with prolonged sepsis: An observational study. J Crit Care 2019 Jun;51:139-144 [FREE Full text] [doi: 10.1016/j.jcrc.2019.02.004] [Medline: 30825787]
29. Lu J, Xun Y, Yu X, Liu Z, Cui L, Zhang J, et al. Albumin-globulin ratio: a novel predictor of sepsis after flexible ureteroscopy in patients with solitary proximal ureteral stones. Transl Androl Urol 2020 Oct;9(5):1980-1989 [FREE Full text] [doi: 10.21037/tau-20-823] [Medline: 33209662]
30. Benediktsson S, Frigyesi A, Kander T. Routine coagulation tests on ICU admission are associated with mortality in sepsis: an observational study. Acta Anaesthesiol Scand 2017 Aug 06;61(7):790-796. [doi: 10.1111/aas.12918] [Medline: 28681428]
31. Dempfle CH, Lorenz S, Smolinski M, Wurst M, West S, Houdijk WPM, et al. Utility of activated partial thromboplastin time waveform analysis for identification of sepsis and overt disseminated intravascular coagulation in patients admitted to a surgical intensive care unit. Crit Care Med 2004;32(2):520-524. [doi: 10.1097/01.ccm.0000110678.52863.f3]
32. Shekar B, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data. 2019 Feb Presented at: International Conference on Advanced Computational and Communication Paradigms; Feb 25-28; Gangtok, Sikkim, India p. 1-8. [doi: 10.1109/icaccp.2019.8882943]
33. Vardon-Bounes F, Ruiz S, Gratacap M, Garcia C, Payrastre B, Minville V. Platelets are critical key players in sepsis. Int J Mol Sci 2019 Jul 16;20(14):3494 [FREE Full text] [doi: 10.3390/ijms20143494] [Medline: 31315248]
34. Sônego F, Castanheira FVES, Ferreira RG, Kanashiro A, Leite CAVG, Nascimento DC, et al. Paradoxical roles of the neutrophil in sepsis: protective and deleterious. Front Immunol 2016 Apr 26;7:155 [FREE Full text] [doi: 10.3389/fimmu.2016.00155] [Medline: 27199981]
35. Korean Liver Cancer Association K, National Cancer Center N. 2018 Korean Liver Cancer Association–National Cancer Center Korea Practice Guidelines for the Management of Hepatocellular Carcinoma. Gut Liver 2019 May 15;13(3):227-299. [doi: 10.5009/gnl19024]
36. Zhang R, Meng J, Lian Q, Chen X, Bauman B, Chu H, et al. Prescription opioids are associated with higher mortality in patients diagnosed with sepsis: A retrospective cohort study using electronic health records. PLoS One 2018 Jan 2;13(1):e0190362 [FREE Full text] [doi: 10.1371/journal.pone.0190362] [Medline: 29293575]

Abbreviations

A/G: albumin/globulin
ANN: artificial neural network
APTT: activated partial thromboplastin time
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic
CDW: Clinical Data Warehouse
CNN: convolutional neural network
EHR: electronic health record
ER: emergency room

XSL•FO
RenderX

FP-growth: frequent pattern growth
ICU: intensive care unit
KISTI: Korea Institute of Science and Technology Information
LR: logistic regression
NRF: National Research Foundation
PT: prothrombin time
ResNet10: residual convolutional neural networks
RF: random forest
RNN-LSTM: long short-term memory recurrent neural networks
Sepsis-3: Third International Consensus Definitions for Sepsis and Septic Shock
SHAP: Shapley Additive Explanations
SMC: Samsung Medical Center
SOFA: Sequential Organ Failure Assessment
XAI: Explainable Artificial Intelligence

Edited by G Eysenbach; submitted 03.03.22; peer-reviewed by S Molani, H Park; comments to author 24.03.22; revised version received 18.04.22; accepted 17.05.22; published 15.06.22

Please cite as:
Yang D, Kim J, Yoo J, Cha WC, Paik H
Identifying the Risk of Sepsis in Patients With Cancer Using Digital Health Care Records: Machine Learning–Based Approach
JMIR Med Inform 2022;10(6):e37689
URL: https://medinform.jmir.org/2022/6/e37689
doi: 10.2196/37689
PMID:

©Donghun Yang, Jimin Kim, Junsang Yoo, Won Chul Cha, Hyojung Paik. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.06.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX