Journal of Biomedical Informatics 109 (2020) 103540

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults

Garrett Eickelberg [a], L. Nelson Sanchez-Pinto [a,b,*], Yuan Luo [a,*]

[a] Department of Preventitive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, 750 N Lake Shore, Chicago, IL 60611, USA
[b] Department of Pediatrics (Critical Care), 225 E. Chicago Avenue, Chicago, IL 60611, USA

ARTICLE INFO

Keywords:
Critical care
Prediction models
Antibiotic stewardship
Machine learning
MIMIC
Electronic health records

ABSTRACT

Unnecessary antibiotic regimens in the intensive care unit (ICU) are associated with adverse patient outcomes and antimicrobial resistance. Bacterial infections (BI) are both common and deadly in ICUs, and as a result, patients with a suspected BI are routinely started on broad-spectrum antibiotics prior to having confirmatory microbiologic culture results or when an occult BI is suspected, a practice known as empiric antibiotic therapy (EAT). However, EAT guidelines lack consensus and existing methods to quantify patient-level BI risk rely largely on clinical judgement and inaccurate biomarkers or expensive diagnostic tests. As a consequence, patients with low risk of BI often are continued on EAT, exposing them to unnecessary side effects. Augmenting current intuition-based practices with data-driven predictions of BI risk could help inform clinical decisions to shorten the duration of unnecessary EAT and improve patient outcomes. We propose a novel framework to identify ICU patients with low risk of BI as candidates for earlier EAT discontinuation. For this study, patients suspected of having a community-acquired BI were identified in the Medical Information Mart for Intensive Care III (MIMIC-III) dataset and categorized based on microbiologic culture results and EAT duration. Using structured longitudinal data collected up to 24-, 48-, and 72-hours after starting EAT, our best models identified patients at low risk of BI with AUROCs up to 0.8 and negative predictive values >93%. Overall, these results demonstrate the feasibility of forecasting BI risk in a critical care setting using patient features found in the electronic health record and call for more extensive research in this promising, yet relatively understudied, area.

1. Introduction

Antibiotics can be life-saving for critically ill patients with bacterial infections (BIs), however, overuse or unnecessary administration can contribute to antimicrobial resistance (AMR) and antibiotic-associated morbidity [1–7]. This is a critical issue, as patients with AMR infections suffer longer hospital stays, treatment complications, higher healthcare costs, and are more likely to die [8–11]. Furthermore, antibiotics can cause harm through gut microbiome dysbiosis, mitochondrial toxicity, and immune cell dysfunction [1–7]. Although clinicians have become more aware of the side effects of antibiotics, it is estimated that up to 50% of antibiotic prescriptions in acute care hospitals in the United States are still either inappropriate or unnecessary [12–17]. Reducing both the amount and duration of unnecessary antibiotic treatments is a commonly proposed strategy to reduce the risk of antibiotic-related side effects [12–15,18]. This is particularly relevant in the intensive care unit (ICU), where concerns for bacterial infections (BI) are high and prescribing antibiotics empirically—prior to having confirmatory bacterial culture results or when an occult BI is suspected—is a common practice [19,20].

Approximately 30–50% of all ICU patients are diagnosed with a BI and their mortality rates can reach as high as 60% in severe infections [20–23]. As a result, providers in the ICU often have a low threshold to start empiric antibiotic therapy (EAT) despite the ramifications of excessive antibiotic use for patients at low risk of BI. Unfortunately, there is no uniform consensus on the appropriate duration of EAT. As a result, clinicians must continually weigh the risks of failing to treat a serious BI against the risks of prescribing inappropriate antibiotic regimens. Moreover, physicians lack objective criteria to identify low BI risk in patients receiving EAT, and rely on clinical intuition and imprecise

Abbreviations: ICU, intensive care unit; BI, bacterial infection; EAT, empiric antibiotic therapy; EHR, electronic health record; MIMIC-III, Medical Information Mart for Intensive Care III dataset.

* Corresponding authors.
E-mail addresses: lazaro.sanchez-pinto@northwestern.edu (L.N. Sanchez-Pinto), yuan.luo@northwestern.edu (Y. Luo).

https://doi.org/10.1016/j.jbi.2020.103540
Received 13 March 2020; Received in revised form 17 July 2020; Accepted 12 August 2020
Available online 16 August 2020

guidelines to balance EAT decisions [3,24–26]. Strategies that shorten unnecessary antibiotic duration in ICU patients when BIs are no longer suspected offer a way to improve patient outcomes, and have been identified as a priority by the Society of Critical Care Medicine as part of their "less is more" campaign [27].

Leveraging electronic health record (EHR) data with machine learning techniques presents an opportunity to accurately identify patients with low risk of BI. The widespread adoption of EHR systems offers investigators access to massive repositories of data generated through routine clinical care and provides opportunities to develop novel prediction algorithms to aid in clinical decision making.

The primary objective of this study was to develop a novel framework to identify ICU patients with a low risk of BI as candidates for earlier EAT discontinuation. The feasibility of this approach was investigated in patients suspected of having a BI by modeling data collected for up to 24-, 48- or 72-hours following the first dose of antibiotics. We compare prediction performance across different model types, data collection windows, and prediction thresholds. The developed algorithm could be used to identify patients at low risk of BI early in their hospitalization who may benefit from early discontinuation of EAT. Furthermore, our EHR-based phenotype of patients suspected of having a BI could be generalized to other datasets and used for additional analyses on antibiotic usage and BI in the ICU.

The detailed data dictionary, code, results been made available at: https://github.com/geickelb/mimiciii-antibiotics-opensource.

2. Materials & methods

2.1. Dataset

A summary of our data extraction and analysis workflow is presented in Fig. 1. The data used in this study was retrieved from the Medical Information Mart for Intensive Care III (MIMIC-III). The MIMIC-III database is an open and de-identified database comprised of health-related data from over 40,000 ICU patients who received care at Beth Israel Deaconess Medical Center between 2001 and 2012 [28,29]. MIMIC-III includes a variety of data such as administrative, clinical and physiological types, which are organized, formatted, processed and de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines [28,29].

2.2. Cohort

Adult patients who were suspected of having a BI upon admission to the ICU were eligible for our study. To match this phenotype, a patient must have: (1) received at least one dose of antibiotics within 96h following ICU admission and (2) had a microbiologic culture within 24h of their first antibiotic dose (Fig. 2). Microbiologic cultures were defined as cultures obtained from any of the following: blood, joint, urine, cerebral spinal fluid (CSF), pleural cavity, peritoneum, or bronchoalveolar lavage. Patients with multiple ICU encounters that met study inclusion criteria were analyzed independently; however, each patient's ICU encounters were assigned to the same train/test split (see *Modeling*).

Antibiotics prescriptions were recorded as the administration of any "antibacterial for systemic use" represented by Anatomical Therapeutic Chemical (ATC) code J01. ATC codes were obtained by first converting national drug codes (NDC) into RxNorm concept unique identifier (RXCUI) codes, and then into ATC codes. Regular expressions were used on prescription names to further filter out erroneous entries and those with missing NDC/RXCUI codes. We calculated the maximum length of cumulative antibiotic days following a microbiologic culture for each ICU encounter. Prescription information in the MIMIC-III database was stored with date level resolution. To accommodate this, the time of each patient's first antibiotic dose (t_0) meeting the phenotype criteria was set to 0:00:00.

Patients were allocated to one of three BI groups: serious BI, non-serious BI/no BI, and unknown BI status (Fig. 3). Given the common occurrence of occult bacterial infections, a direct inference of BI status could not be made based off of microbiological culture results alone. Therefore, patient's BI statuses were assigned based both on their microbiologic culture results (positive vs. negative) and duration of their antibiotic treatment (short [≤96h] vs. prolonged [>96h]). In this paradigm, patients with positive microbiologic culture and prolonged antibiotic treatment were considered to have serious BIs (prediction

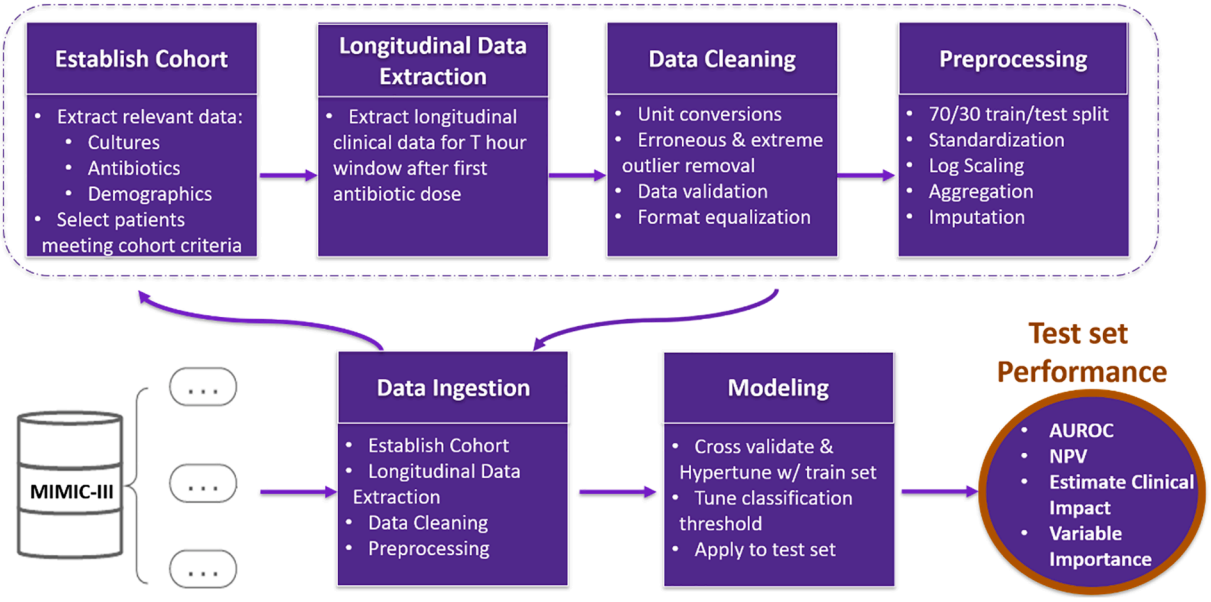

Fig. 1. Data Ingestion and Analysis Framework Overview. Raw data is ingested from the MIMIC-III database. First a cohort of adult patients suspected of having SBI is established, and both longitudinal and categorical data is extracted over the T = 24-,48-, or 72-hour window following their first antibiotic dose that corresponds with an microbiologic culture. Next, data is cleaned, formatted, and preprocessed prior to modeling. The cohort is then filtered to patients with positive microbiologic culture and prolonged antibiotics, and microbiologic culture negative with short antibiotics. A 70/30 train/test set split is then applied. Scaling and standardization are performed on each set independently. Missing values were imputed using median values from the training set. Machine learning models are hypertuned on the training set and applied to the test set. Finally, classification thresholds are tuned, and model performance metrics are output.

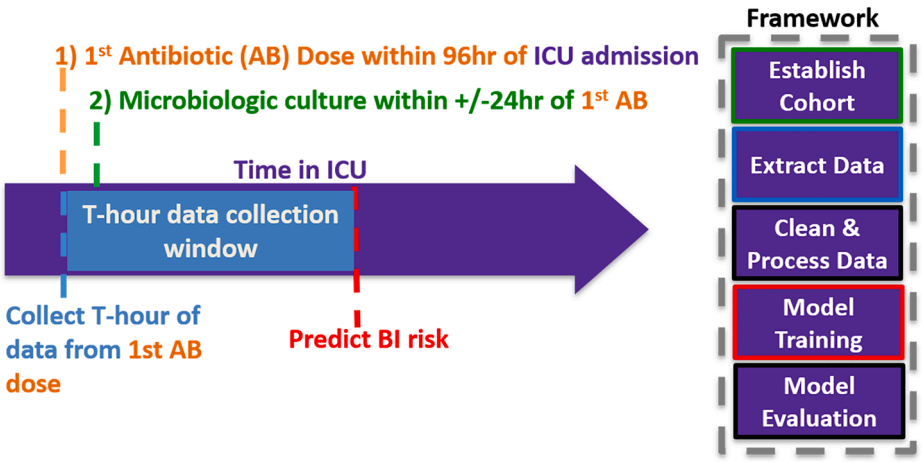

Fig. 2. Phenotype Criteria for BI Suspicion at ICU Admission. A patient's first Antibiotic (AB) dose (t_0) needs to: (1) be administered within 96h following ICU admission and (2) have an microbiologic culture within 24h and (1) be administered within 96h following ICU admission. Clinical Data is collected for up to T = 24-, 48-, or 72-hours after first antibiotic dose.

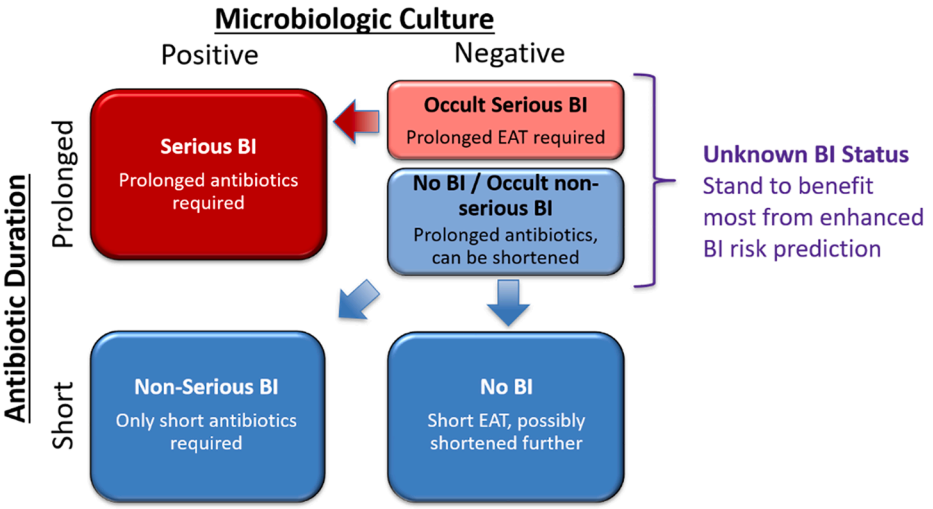

Fig. 3. Classification of BI status and framing of the clinical prediction problem. Patient BI status can be classified into three groups based on duration of antibiotics and microbiological results: "Serious BI" are those with positive microbiological cultures receiving antibiotics for >96h and are the cases in model training. "Non-serious BI" and "No BI" patients are those with antibiotics ≤96h and are the controls in the model training. "Unknown BI status" are patient who received empiric antibiotic therapy [EAT] for >96h despite negative microbiological cultures, and are the group of patients most likely to benefit from correct BI risk prediction. The unknown BI status group may be conceptually divided into patients with "occult serious BI" who are likely more similar to the cases than to the controls, and patients with "no BI or occult non-serious BI" who are likely more similar to the controls than to the cases.

events), whereas those with negative cultures and short antibiotic treatment were considered to have no BIs (prediction non-events). Additionally, patients with short antibiotic treatment and positive microbiologic culture were considered non-serious BIs. Due to the possibility of occult infections, patients who received prolonged antibiotics despite having a negative microbiologic culture had less clear infection statuses, and were thus coded as unknown BI status. Conceptually, patients in that group could be further divided into those with an occult serious BI and those with either no BI or an occult non-serious BI. These patients were separated from the dataset prior to model training and testing, and were later used assess the clinical utility of the prediction model by testing its ability to identify patients at low-risk BI in that population.

To control for *Staphylococcus* culture contamination, we required two consecutive *Staphylococcus* positive cultures to be considered microbiologic culture positive. Additionally, we coded patients that died within 24h of their last antibiotic dose as prolonged antibiotic treatment (n = 1266). To accommodate for date-level resolution on prescription timings, we utilized a conservative 96h threshold for short vs prolonged antibiotic duration.

2.3. Data extraction

We extracted static and longitudinal patient clinical data from the MIMIC-III database using open source code provided by the MIMIC-III team (Table 1). Longitudinal data was restricted to either the T = 24-, 48-, or 72-hour cutoff following the administration date of the first antibiotic dose (t_0:t_{0+T}) (Fig. 2).

2.4. Cleaning & pre-processing

The raw clinical data extracted for the purpose of this study were first cleaned and formatted to address data quality issues and then pre-processed to facilitate usability by selected machine learning models. The first cleaning step was to address desperate units of measurement by converting each variable into designated units (Table 1). Next, conservative thresholds were set to review erroneous values and data entry errors for removal based upon a combination of reference laboratory value limits, clinical knowledge, three sigma outlier criteria, and manual audit of a subset of free-text to confirm concordance. Finally, event and windowed continuous variables, such as administration of renal replacement therapy or mechanical ventilation were coded and discretized. The cleaned data were then converted into unit variances following as in (1), where $X_{(-/short)}$ is the median value of the patients

Table 1
Extracted Data- Raw variables and units extracted from the corresponding table in the MIMIC-III database.

MIMIC-III TABLE	Data Collected	Unit	% Missingness (T = 24–72h)
Diagnoses	ICD-9 codes (Elixhauser Comorbidity Index)	categorical	0–0
Admissions	Age	years	0–0
	Ethnicity	categorical	0–0
	Gender	categorical	0–0
ChartEvents	Blood pressure (systolic, diastolic)	mmHg	0.2–0
	Glasgow Coma Scale	GCS score	72.5–53.3
	Glucose	mg/dL	0.5–0.1
	Heart rate	bpm	0–0
	Peripheral oxygen Saturation (SpO2)	%	0–0
	Temperature	deg. C	1.6–0.2
	Ventilation status	categorical	1.3–0.9
	Weight	kg	8.4–8.4
InputEvents	Dobutamine	mcg/kg/min	98.8–98.3
	Dopamine	mcg/kg/min	94.9–94
	Epinephrine	mcg/kg/min	97.9–97.6
	Norepinephrine	mcg/kg/min	83.1–80.1
	Phenylephrine	mcg/kg/min	86.2–83.2
	Renal replacement therapy	pos/neg	0–0
	Vasopressin	mcg/kg/min	98–97
LabEvents	Bands	%	87.3–82.6
	Serum bicarbonate	mEq/L	2.1–0.3
	Bilirubin	mg/dL	60.1–47.8
	Blood urea nitrogen (BUN)	mg/dL	2–0.3
	Serum chloride	mEq/L	1.9–0.3
	Serum creatinine	mg/dL	2–0.3
	Serum glucose	mg/dL	0.5–0.1
	Hemoglobin	g/dL	2.6–0.3
	International Normalized Ratio (INR)	ratio	24.9–13.3
	Serum lactate	mmol/L	48.1–42.3
	Urine leukocyte	pos/neg	69.5–57.6
	Urine nitrite	pos/neg	69.5–57.6
	Partial pressure of arterial oxygen (PaO2)/fraction of inspired oxygen (FiO2) ratio	ratio	67.9–65.1
	Partial thromboplastin time (PTT)	sec	25.2–13.8
	Partial pressure of arterial carbon dioxide (pCO2)	mmHg	39.9–34
	Serum pH	n/a	41.9–36.7
	Platelet count	K/uL	2.6–0.3
	Serum potassium	mEq/L	1.6–0.3
	White blood cell count	K/uL	2.9–0.3
	Serum calcium	mmol/L	63.1–56.6

with negative microbiologic culture and short duration EAT. Next, longitudinal and ordinal clinical variables spanning t_0:t_{0+T} were aggregated to produce single value(s) for each parameter using the operation that conferred the highest likelihood of infection (minimum, maximum or both). Lastly, categorical variables were encoded to dummy variables using the one-hot-encoding technique. The final dataset was represented by a 52-dimension feature vector.

$$Z = \frac{X - X_{(-/short)}}{IQR_{(-/short)}} \quad (1)$$

2.5. Modeling

The patients with positive microbiological cultures and prolonged antibiotic duration (serious BI) and those with short antibiotic duration (no BI or non-serious BI) were split into a training and test set following a 70/30 split based upon unique ICU stay identifiers. Cohort splitting was performed on unique ICU stay identifiers where individual patients were sequestered to either the training or testing set to prevent testing set contamination. We chose to impute missing values with median values from the training set in order to facilitate implementation into a clinical setting. Empirical studies have suggested that including imputed values with high missingness can improve model clinical utility, so we chose to include imputed values with high missingness in our model (Table 1) [30,31].

The final dataset was modeled using a variety of machine learning algorithms, including Ridge regression [32], Random Forests [33], support vector classifier (SVC) [34], extreme Gradient Boosted decision Tree (XG Boost) [35], K-Nearest Neighbors (K-NN), and Multilayer Perceptron (MLP). These models were chosen using a set of criteria that included each model's relative interpretability, approach to handling nonlinearity, and ability to model categorical and continuous features. A soft voting classifier, or ensemble of all other models, was also used to test for significant performance gains or losses.

Class imbalance was addressed by classification threshold tuning and modeling specific class balancing parameters, such as bootstrapping and class weights, during hyperparameter tuning in order to simplify the modeling workflow. Modeling hyperparameters were tuned using 10-fold cross validation with a binary cross entropy loss function on the training set. The binary classification threshold was tuned in 10-fold cross validation to achieve a high sensitivity (sensitivity $\geq$ 0.9) and was averaged across all folds. This high sensitivity was chosen in order to reduce the number of false negatives and predict low BI risk with higher certainty. Threshold tuned model performances were assessed on the test set using area under the receiver operator curve (AUC), F1 score, negative predictive value (NPV), precision, and recall.

3. Results

3.1. Cohort

We identified a total of 19,633 ICU encounters (15,412 unique patients) in the MIMIC-III data that met inclusion criteria for our study. Within this set, we filtered our cohort down to 12,232 ICU encounters (10,290 unique patients) that had either prolonged antibiotics and positive microbiologic culture, or short antibiotics and negative microbiologic culture (Table 2). Table 3 summarizes the breakdown of these patients across the train/test splits. Additionally, 7401 ICU encounters (6520 unique patients) with unknown BI status (prolonged antibiotics and negative microbiologic culture) were set aside to test the prediction model's ability to identify patients at low risk BI in that population.

Table 4 summarizes the test set results for each threshold tuned model. The performance across the models for each T-hour test set showed little variation, where XGBoost and Random Forests slightly outperformed the other models in terms of AUC, F1 score, NPV, and precision. As the data window was increased from 24 to 72h, there were

Table 2
Demographics- distribution of cohort demographics.

Variable	Mean/stdev
Gender- N, %	
Female	5709 (47%)
Male	6523 (53%)
Age (yr)	64.7 ± 17.0
Ethnicity- N, %	
African-American	1385 (11%)
White	8855 (72%)
Hispanic	507 (4%)
Other	1485 (12%)

Table 3
Cohort Split- Breakdown of the train/test split against patient classes.

Microbiologic Culture	Antibiotic Duration[a]	BI Status Classification	Train No. (%)	Test No. (%)	Total No. (%)
Negative	Short	Positive	5512 (65%)	2355 (65%)	7867 (65%)
Positive	Prolonged	Negative	1693 (20%)	745 (20%)	2438 (20%)
Positive	Short	Negative	1296 (15%)	631 (15%)	1927 (15%)
Negative	Prolonged	Unknown	N/A	N/A	7401 (100%)

[a] Time on antibiotics, short (≤96h) vs. prolonged (>96h).

Table 4
Preliminary Model Results - Modeling parameters for each model on the test set using the high sensitivity threshold.

Model	AUC	F1	NPV	Precision	Recall	High Sensitivity Threshold
72-hour Test set						
Random Forests Classifier	0.793	0.431	0.941	0.284	0.891	0.124
XGBoost	0.795	0.439	0.943	0.291	0.891	0.096
MLP Classifier	0.779	0.395	0.948	0.25	0.936	0.09
Logistic Regression	0.781	0.423	0.932	0.278	0.876	0.298
SVC	0.778	0.425	0.935	0.28	0.881	0.101
K-NN	0.734	0.357	0.936	0.219	0.963	0.04
Voting Classifier	0.793	0.429	0.946	0.281	0.905	0.147
48-hour Test set						
Random Forests Classifier	0.788	0.43	0.943	0.283	0.897	0.126
XGBoost	0.796	0.436	0.946	0.288	0.9	0.091
MLP Classifier	0.771	0.456	0.92	0.318	0.805	0.084
Logistic Regression	0.774	0.421	0.938	0.275	0.893	0.296
SVC	0.773	0.42	0.941	0.274	0.9	0.099
K-NN	0.733	0.393	0.922	0.252	0.887	0.044
Voting Classifier	0.788	0.436	0.939	0.29	0.881	0.147
24-hour Test set						
Random Forests Classifier	0.774	0.424	0.944	0.277	0.905	0.258
XGBoost	0.776	0.416	0.94	0.271	0.901	0.104
MLP Classifier	0.764	0.439	0.925	0.297	0.84	0.087
Logistic Regression	0.764	0.411	0.94	0.266	0.907	0.302
SVC	0.763	0.411	0.937	0.267	0.9	0.105
K-NN	0.714	0.382	0.922	0.243	0.903	0.044
Voting Classifier	0.776	0.421	0.939	0.275	0.895	0.177

small increases in AUC across the best performing models for each time window. Fig. 4 summarizes the ROC curve for all the T = 24-hour models where all, except K-nearest neighbors, performed similarly. Additionally, when tested with the 72-hour test data, the 24-hour Random Forests model obtained an AUC of 0.787 (~0.013 increase). Similarly, the 72-hour Random Forests model produced an AUC of 0.765 (~0.028 decrease) when tested on the 24-hour data. These changes in AUC suggest that both the 24-hour and 72-hour models maintain similar model performances when making predictions on data collected over 48-hour longer and shorter collection windows, respectively.

Fig. 5 displays how variable importance changed across the models. For this plot, a list of 20 variables was selected based on the top ten most important variables for the Random Forests, logistic regression, XGBoost, and SVC models. Fig. 5 suggests that although the models perform similarly, each model prioritized predictors. This interpretation is reinforced by the results of the soft voting ensemble models, which performed comparably to the best performing model within each T-hour test set. This further suggests that the models are identifying the same or similar patients regardless of the underlying algorithm.

The T = 24-hour Random Forests model was chosen for the subsequent analyses given that the T = 24-hour timepoint provides more clinical utility, and thus the Random Forests model was the best performing model within this timepoint. Table 5 summarizes the confusion matrix for this model with a high sensitivity classification threshold (0.26) in the test set. The model achieved an NPV of 0.944 in the test set; however, this figure is based on the 0.20 BI prevalence from the training and testing set. Fig. 6 displays how the model NPV changes as a function of population BI prevalence and classification threshold. We found that as the BI prevalence changed from 0.5 to 0.1, the NPV of the T = 24-hour random forests model changed from 0.82 to 0.98 when using a high sensitivity threshold, and 0.59 to 0.93 when using a 0.5 threshold. These results suggest that our model performance will be more robust to changes in prevalence when using the high sensitivity prediction threshold. The remaining patients falsely predicted as negatives by all of the T = 24-hour models were investigated for observable patterns. These investigations suggested that the false negatives are a heterogeneous group with no reproducible patterns.

Of the 1208 true negatives, 458 (37.9%) cases received antibiotics for 24h or less, while 750 (62.1%) received antibiotics greater than 24h. We estimated that 1289 out of the 2375(53.2%) total antibiotic days administered to patients in the true negative group could have been avoided if our model with a high sensitivity threshold were hypothetically used to stop EAT early.

3.2. Performance in patient set with unknown BI

Finally, the best performing T = 24-hour Random Forests model was applied to the patient group with unknown BI status, which are those who stand to benefit the most from correct BI risk prediction. Using the high sensitivity and 0.5 probability thresholds, the model predicted 861 out of 7,401 (11.6%) and 5,525 out of 7,401 (74.7%) patients to be at low risk of BI, respectively (Table 6). Using the NPV from the test set with high sensitivity and 0.5 thresholds (NPV = 94.5%, 84.3%) we estimated that approximately 48 (0.6%) and 860 (11.6%) of all unknown BI status patients would have been predicted to have a low BI risk but actually have had a BI (false negatives). By subtracting these estimated false negative patients from the total negative predictions, we estimated that the high sensitivity and 0.5 thresholds would have theoretically benefited 813 (11.0%) and 4,664 (63.0%) patients, respectively. We estimated that as a lower bound, our T = 24-hour Random Forests model with a high sensitivity threshold could have reduced approximately 5,684 (9.5%) antibiotic days administered to this group, and as an upper bound with the 0.5 probability threshold could have reduced approximately 35,831 (60.0%) antibiotic days. A manual chart review and clinical assessment of 10 patient records with unknown BI status (5 predicted high BI risk, 5 predicted low BI risk) found that 8 out 10 model BI risk classifications matched the clinical reviewer's assessment of BI risk, 2 out of 10 were probably correct but remained indeterminate, and 0 out of 10 were misclassified (Appendix A).

4. Discussion

In this study, we developed a novel framework to extract patient features from raw clinical data and identify patients at low risk of BI who, in theory, could benefit from earlier EAT discontinuation within

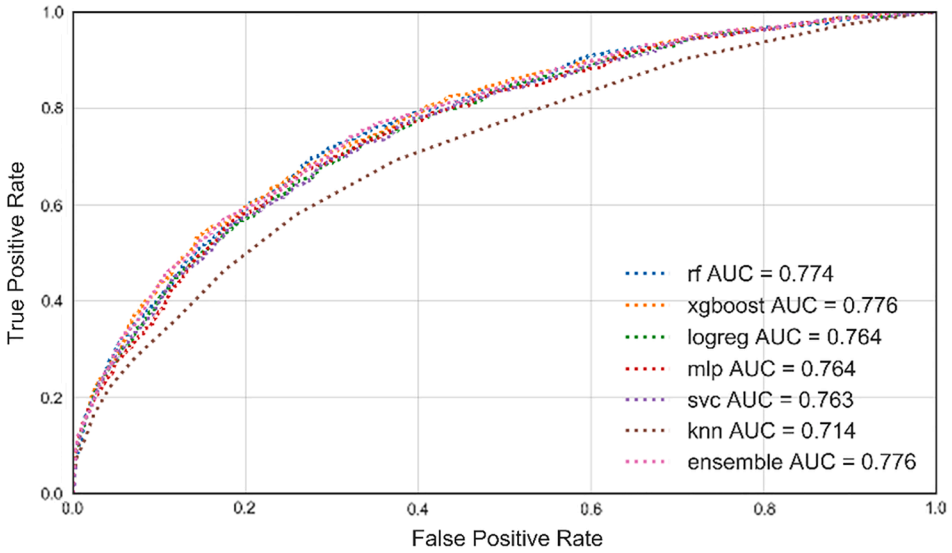

Fig. 4. Receiver operating characteristic curves for all T = 24-hour models. We use different colors and line styles to differentiate models. AUC: Area under the curve.

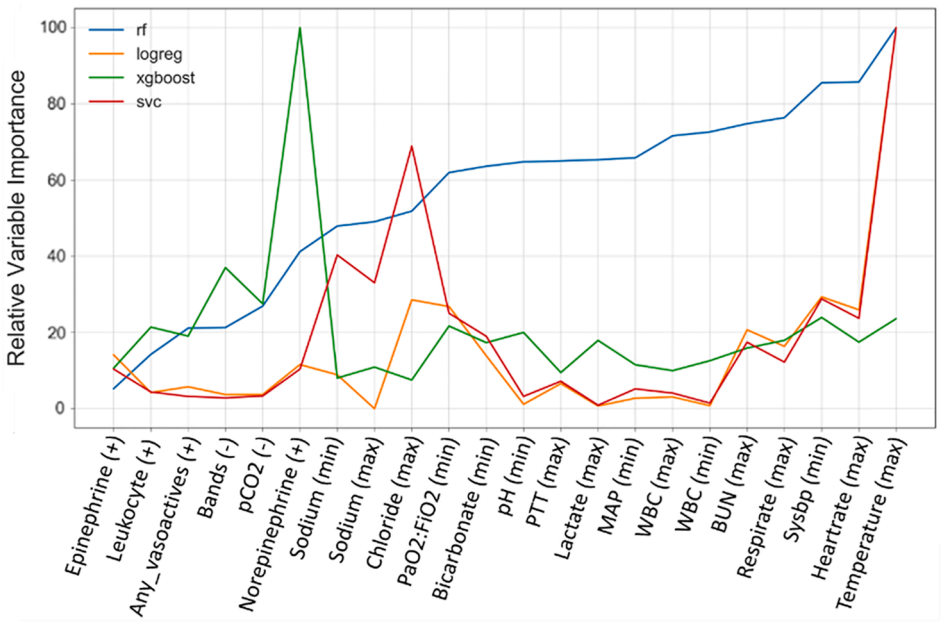

Fig. 5. Stacked Relative Variable Importance Across Prediction Models. Variable importance for Random Forests and XGBoost were based on standardized Gini importance, while SVC and logistic regression used standardized coefficients. Variable importance values from all models were scaled relative to the value of the most important variable for all 20 values in the variable list. pCO_2: carbon dioxide partial pressure; PaO2:FiO2: ratio of arterial oxygen partial pressure to fractional inspired oxygen; PTT: platelets; MAP: average arterial blood pressure over one cardiac cycle; WBC: white blood cell count; BUN: Blood urea nitrogen; sysBP: Systolic blood pressure.

Table 5
Confusion Matrix Statistics- Test set classification summary for the T = 24-hour Random Forests model with a high sensitivity threshold.

	True Negatives (%)	False Positives (%)	False Negatives (%)	True Positives (%)
High Sensitivity Threshold	1208 (32.4%)	1773 (47.5%)	71 (1.9%)	679 (18.2%)
0.5 Probability Threshold	2826 (75.7%)	155 (4.2%)	521 (14.0%)	229 (6.1%)

24h of initiation. Our main finding is that our models can predict patients with low risk of BI with good performance when applied to structured clinical data collected for T = 24-hours after the EAT initiation. We also found that increasing the data collection time and model complexity yielded only slight performance increases. Finally, our results suggest by that applying our T = 24-hour Random Forests model with a high sensitivity threshold to the patient set with unknown BI status (prolonged antibiotics and negative microbiologic culture), we would be able to identify around 11.6% of patients as candidates for EAT removal with high confidence and could reduce total antibiotic days by approximately 9.5%.

Designing data-driven approaches to accurately stratify patients based on their BI risk has the potential to greatly improve antibiotic stewardship efforts. Antibiotic stewardship in the ICU can be viewed as a two-stage process. The first stage requires administering broad-spectrum antibiotics to maximize treatment of serious BI. In the second stage, physicians either stop EAT for patients at low risk of BI or narrow the spectrum of antibiotics once the infection is characterized [3]. Many stewardship techniques focusing on the later stage hinge upon sensitive and specific identification and monitoring of BI risk. Bacterial cultures and inflammatory biomarkers are currently the most common methods of monitoring BI risk in the ICU, but are not necessarily optimal. Bacterial cultures, the current gold standard for diagnosing BI, may take days to result and are often unreliable in detecting all BIs [36]. To

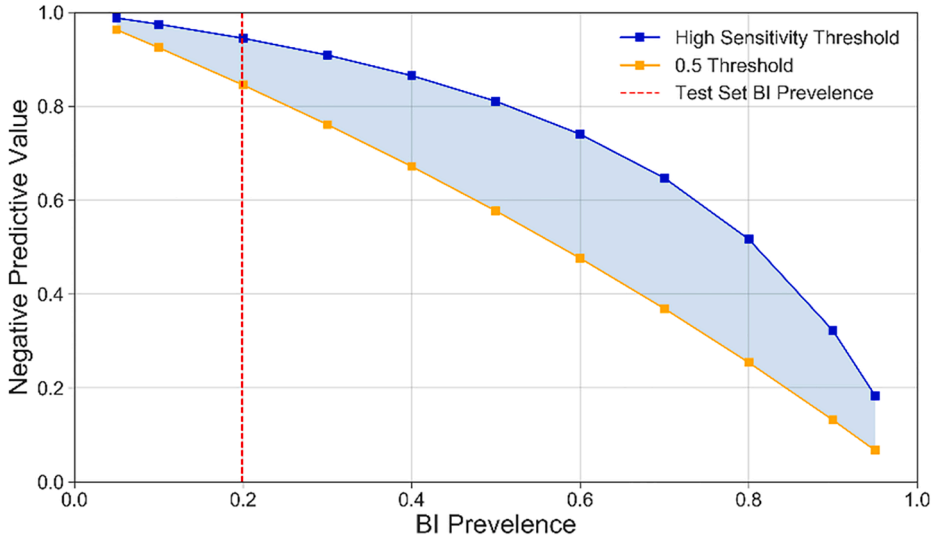

Fig. 6. NPV across BI prevalence for T = 24-hour Random Forests tuned and 0.5 prediction thresholds. NPV was simulated for a variety of BI prevalence values using the sensitivity and 1-specficity for the high sensitivity and 0.5 prediction thresholds from the test set.

Table 6
Prolonged antibiotic negative microbiologic culture predictions- prediction distribution for the T = 24-hour Random Forests model.

	High Sensitivity Threshold	0.5 Threshold
Predicted low BI risk	861 (11.6%)	5525 (74.7%)
Predicted high BI risk	6540 (88.4%)	1876 (25.3%)

address this, bacterial cultures are frequently supplemented with Gram staining, which provide additional information more immediately about a patient's BI risk. However, Gram staining suffers from high variability and low reliability that results from individual differences in slide preparation and interpretation [37–39]. Assays based on inflammatory biomarkers, such as C-reactive protein and procalcitonin, have improved sensitivity and specificity for detecting community-acquired infections, but have high rates of false-positives and -negatives for hospital-acquired infections [3,40–42]. Newer rapid multiplex diagnostics for infectious organisms have also been introduced; however, these are still being tested for efficacy, costly, and not yet widely available [43]. Designing better methods to identify patients with low risk for BI is critical to shorten the duration of unnecessary EAT and facilitate antibiotic stewardship.

Numerous prior studies have presented EHR-based machine learning models and clinical decision support systems to predict infection related conditions, such as bacteremia, sepsis, and ICU mortality [44–55]. The goal of such models has been to ensure all septic and/or bacteremic patients are identified and treated early with appropriate antibiotic regimens [45–48,51–53,55]. For instance, Nemati et al. achieved AUROCs ranging from 0.83 to 0.85 in predicting the early onset of sepsis using data collected during the 12, 8, 6, and 4 h prior to diagnosis for patients across two Emory University hospitals and the MIMIC-III dataset. In contrast to these prior studies, the models we present differ by clinical timeframe (it is intended to be used after a patient is already suspected of having BI and has started EAT) and by the goal of the model (identify patients on EAT who are candidates for EAT discontinuation). Currently, no other prominent EHR-based prediction models exist with the goal of identifying patients on EAT with low risk of having BI who are candidates for EAT discontinuation. Existing methods for forecasting patient-level BI risk have focused around the use of protein and genetic biomarkers [6,40,41]. The models we present rely on data commonly recorded in the ICU and do not require any specialized laboratory diagnostics or data from current BI risk prediction methods. Our study adds to the body of research surrounding EHR-based prediction models and provides a complementary approach to biomarker-based forecasting of patient-level BI risk. When used in combination with current BI risk metrics and clinical intuition, our model promises to help assist care providers in the de-escalation process of antibiotic stewardship.

For our clinical use case, false negative patients, i.e. those with a serious BI who were predicted as unlikely to have an infection, encompass the largest source of potential patient harm given the risk of untreated BIs in the ICU and therefore need to be minimized. Similarly, the largest source of potential patient benefit of our model from the current standard of care comes from reducing the number of antibiotic days given to patients who don't have known BI. Our T = 24-hour Random Forests model uses a high sensitivity decision threshold to in order to reduce false negative predictions and therefore improve the potential clinical utility in an ICU setting.

We recognize several limitations of this study. First, the retrospective data used was collected for clinical care purposes at a single academic medical center. The retrospective design of our study required us to infer information regarding BI suspicion, consecutive antibiotic days, and culture results based upon sensible criteria that may not completely reflect real world conditions. To address this, chart review and a variety of other quality checks were performed throughout the workflow to ensure appropriate coding of outcomes. Results from our 10-patient chart review of unknown BI status patients found two indeterminate cases and zero misclassifications by our proposed model. Details in the chart notes of one of these indeterminate cases suggested that this patient experienced a prolonged stay in the emergency department prior to transferring to the ICU and that the data from the emergency department was not available in the MIMIC-III dataset. This case suggests that the performance of our phenotype and model can be improved with more complete data on patients prior to ICU transfer. Future work will include retrospective data from additional ICU centers for external model validation and assessment of clinical utility, including data prior to ICU admission. Next, our estimates of antibiotic reduction provide an upper and lower bound on the potential clinical impacts of our model and makes numerous assumptions. To better understand the clinical utility of our model, further study is necessary to test the hypothesis that discontinuing antibiotic therapy on the patients predicted as low risk of BI would clinically benefit them. In future work, we will perform a propensity-matched analysis to estimate the effects of receiving short vs. prolonged antibiotics on outcome in patients with a predicted low risk of BI. Finally, the longitudinal patient data collected over T = 24-,48-, or 72-hours was aggregated prior to modeling using the aggregation

function(s) most associated with increased BI risk for each variable. With this design, the time for patients to exhibit symptoms most indicative of BI risk increases as the data collection window increases; however, time-window aggregation methods do capture temporal patterns present in the data to the fullest extent. To better leverage the longitudinal nature of our data, future work will focus on testing more complex algorithms to explore temporal trends and improve model performances.

5. Conclusion

The goal of this paper was to detail the design and initial application of a novel collection of algorithms which extract patient features from clinical data and identify patients at low risk of BI who can be safely removed from EAT at 24-hours after initiation. Our models achieved up to 0.8 AUC and demonstrate the feasibility of forecasting BI risk in a critical care setting using patient features found in the EHR. Future work will focus on validating models with external datasets, measuring clinical utility more accurately, and improving model performance by accounting for temporal information in patient data. Overall, these results call for more extensive research in this promising, yet relatively understudied, area.

CRediT authorship contribution statement

Garrett Eickelberg: Methodology, Software, Validation, Visualization, Investigation, Formal analysis, Writing - original draft, Funding acquisition. **L. Nelson Sanchez-Pinto:** Methodology, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Yuan Luo:** Methodology, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

LNSP and YL are co-corresponding authors. This research is partly supported by grant R21 LM012618 (Luo) from the National Institutes of Health, grant 5T32LM012203-04 from the National Library of Medicine (Eickelberg), and grant R21 HD096402 from the National Institute of Child Health & Human Development (Sanchez-Pinto).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103540.

References

[1] J.A. Claridge, P. Pang, W.H. Leukhardt, J.F. Golob, J.W. Carter, A.M. Fadlalla, Critical analysis of empiric antibiotic utilization: establishing benchmarks, Surg. Infections 11 (2) (2010) 125–131.
[2] M.P. Francino, Antibiotics and the human gut microbiome: dysbioses and accumulation of resistances, Front. Microbiol. 6 (2015) 1543.
[3] C.-E. Luyt, N. Bréchot, J.-L. Trouillet, J. Chastre, Antibiotic stewardship in the intensive care unit, Crit. Care 18 (5) (2014) 480.
[4] Z. Thomas, F. Bandali, J. Sankaranarayanan, T. Reardon, K.M. Olsen, A Multicenter evaluation of prolonged empiric antibiotic therapy in adult ICUs in the United States, Crit. Care Med. 43 (12) (2015) 2527–2534.
[5] C.H. Weiss, S.D. Persell, R.G. Wunderink, D.W. Baker, Empiric antibiotic, mechanical ventilation, and central venous catheter duration as potential factors mediating the effect of a checklist prompting intervention on mortality: an exploratory analysis, BMC Health Services Res. 12 (2012) 198.
[6] G. Zilahi, M.A. McMahon, P. Povoa, I. Martin-Loeches, Duration of antibiotic therapy in the intensive care unit, J. Thoracic Dis. 8 (12) (2016) 3774–3780.
[7] N. Arulkumaran, M. Routledge, S. Schlebusch, J. Lipman, M.A. Conway, Antimicrobial-associated harm in critical care: a narrative review, Intensive Care Med. (2020).
[8] Surveillance of Antimicrobial Resistance in Europe. In: Control ECfDPa, ed2017.
[9] F. Prestinaci, P. Pezzotti, A. Pantosti, Antimicrobial resistance: a global multifaceted phenomenon, Pathog Glob Health. 109 (7) (2015) 309–318.
[10] P. Dadgostar, Antimicrobial resistance: implications and costs, Infection Drug Resistance 12 (2019) 3903–3910.
[11] Antibiotic resistance threats in the United States, in: Prevention CfDCa, Services UDoHaH, eds2013, 2013.
[12] More evidence on link between antibiotic use and antibiotic resistance. ScienceDaily: European Centre for Disease Prevention and Control (ECDC); 07/27/2017, 2017.
[13] Antimicrobial resistance: global report on surveillance, World Health Organization, 2014.
[14] L.J. Shallcross, D.S.C. Davies, Antibiotic overuse: a key driver of antimicrobial resistance, Br. J. Gen. Pract. 64 (629) (2014) 604–605.
[15] C.A. Michael, D. Dominey-Howes, M. Labbate, The antimicrobial resistance crisis: causes, consequences, and management, Front. Public Health 2 (2014) 145.
[16] Core Elements of Hospital Antibiotic Stewardship Programs | Antibiotic Use | CDC, 2019.
[17] B.C. Camins, M.D. King, J.B. Wells, et al., Impact of an antimicrobial utilization program on antimicrobial use at a large teaching hospital: a randomized controlled trial, Infect. Control Hosp. Epidemiol. 30 (10) (2009) 931–938.
[18] B.G. Bell, F. Schellevis, E. Stobberingh, H. Goossens, M. Pringle, A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance, BMC Infect. Dis. 14 (1) (2014) 13.
[19] D.A. Goff, T.M. File, The risk of prescribing antibiotics "just-in-case" there is infection, Seminars Colon Rectal Surg. 29 (1) (2018) 44–48.
[20] J.-L. Vincent, J. Rello, J. Marshall, et al., International study of the prevalence and outcomes of infection in intensive care units, JAMA 302 (21) (2009) 2323–2329.
[21] D.C. Angus, W.T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, M.R. Pinsky, Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care, Crit. Care Med. 29 (7) (2001) 1303–1310.
[22] F.B. Mayr, S. Yende, D.C. Angus, Epidemiology of severe sepsis, Virulence 5 (1) (2014) 4–11.
[23] J.L. Vincent, E. Abraham, D. Annane, G. Bernard, E. Rivers, G. Van den Berghe, Reducing mortality in sepsis: new directions, Crit Care. 6 (Suppl 3) (2002) S1–S18.
[24] K. Andre, M. Mark, K. Michael, et al., Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia, Am. J. Respir. Crit. Care Med. 171 (4) (2005) 388–416.
[25] R.P. Dellinger, M.M. Levy, A. Rhodes, et al., Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012, Crit. Care Med. 41 (2) (2013) 580–637.
[26] J.S. Solomkin, J.E. Mazuski, J.S. Bradley, et al., Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the Surgical Infection Society and the Infectious Diseases Society of America, Clin. iInfect. Dis.: Off. Publ. Infect. Dis. Soc. Am. 50 (2) (2010) 133–164.
[27] J.J. Zimmerman, Society of critical care medicine presidential address—47th annual congress, February 2018, San Antonio, Texas, Crit. Care Med. 46 (6) (2018) 839–842.
[28] A. Kumar, D. Roberts, K.E. Wood, et al., Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock, Crit. Care Med. 34 (6) (2006) 1589–1596.
[29] D. Misquitta, Early Prediction of Antibiotics in Intensive Care Unit Patients [Master's Thesis]: Biomedical Informatics, Harvard Medical School, 2013.
[30] Y. Luo, P. Szolovits, A.S. Dighe, J.M. Baron, 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. (1527-974X (Electronic)).
[31] Y. Luo, P. Szolovits, A.S. Dighe, J.M. Baron, Using Machine Learning to Predict Laboratory Test Results, (1943-7722 (Electronic)).
[32] S. Le Cessie, J.C. Van Houwelingen, Ridge estimators in logistic regression, J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 41 (1) (1992) 191–201.
[33] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[34] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[35] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[36] C. Luna, D. Blanzaco, M. Niederman, et al., Resolution of ventilator-associated pneumonia: Prospective evaluation of the clinical pulmonary infection score as an early clinical predictor of outcome*, Crit. Care Med. 31 (3) (2019) 676–682.
[37] F. Blot, B. Raynard, E. Chachaty, C. Tancrède, S. Antoun, G. Nitenberg, Value of gram stain examination of lower respiratory tract secretions for early diagnosis of nosocomial pneumonia, http://dxdoiorg/101164/ajrccm16259908088, 2000.
[38] M. Campion, G. Scully, Antibiotic use in the intensive care unit: optimization and de-escalation, J. Intensive Care Med. 33 (12) (2018) 647–655.
[39] L.P. Samuel, J.-M. Balada-Llasat, A. Harrington, R. Cavagnolo, Multicenter Multicenter Assessment of Gram Stain Error Rates, 2016.
[40] E. de Jong, J.A. van Oers, A. Beishuizen, et al., Efficacy and safety of procalcitonin guidance in reducing the duration of antibiotic treatment in critically ill patients: a randomised, controlled, open-label trial, Lancet. Infect. Dis 16 (7) (2016) 819–827.
[41] P. Schuetz, Y. Wirz, R. Sager, et al. Procalcitonin to initiate or discontinue antibiotics in acute respiratory tract infections, The Cochrane database of systematic reviews 10 (2017) Cd007498.
[42] J.W. Cals, M.H. Ebell, C-reactive protein: guiding antibiotic prescribing decisions at the point of care, Br. J. Gen. Pract. 68 (668) (2018) 112–113.

[43] J.R. Paonessa, R.D. Shah, C.I. Pickens, et al., Rapid detection of Methicillin-resistant Staphylococcus aureus in BAL: a pilot randomized controlled trial, Chest 155 (5) (2019) 999–1007.
[44] L. Ward, M. Paul, S. Andreassen, Automatic learning of mortality in a CPN model of the systemic inflammatory response syndrome, Math. Biosci. 284 (2017) 12–20.
[45] L. Ward, J.K. Møller, N. Eliakim-Raz, S. Andreassen, Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation, IFAC-PapersOnLine. 51 (27) (2018) 116–121.
[46] J.D. Parente, K. Möller, G.M. Shaw, J.G. Chase, Hidden Markov models for sepsis classification, IFAC-PapersOnLine 51 (27) (2018) 110–115.
[47] M. Paul, S. Andreassen, A.D. Nielsen, et al., Prediction of bacteremia using TREAT, a computerized decision-support system, Clin. Infect. Dis.: Off. Publ. Infect. Dis. Soc. Am. 42 (9) (2006) 1274–1282.
[48] E. Sheetrit, N. Nissim, D. Klimov, Y. Shahar, Temporal probabilistic profiles for sepsis prediction in the ICU, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019; Anchorage, AK, USA.
[49] S.M. Vieira, J.P. Carvalho, A.S. Fialho, S.R. Reti, S.N. Finkelstein, J.M.C. Sousa, A decision support system for ICU readmissions prevention, in: Proceedings of the 2013 Joint Ifsa World Congress and Nafips Annual Meeting (Ifsa/Nafips), 2013, 251–256.
[50] Y. Luo, Y. Xin, R. Joshi, L. Celi, P. Szolovits, Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements, in: Paper presented at: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence; 02/12/2016, 2016.
[51] E. Gultepe, J.P. Green, H. Nguyen, J. Adams, T. Albertson, I. Tagkopoulos, From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system, J. Am. Med. Inform. Assoc. 21 (2014) 315–325.
[52] R. Brause, F. Hamker, J. Paetz, Septic shock diagnosis by neural networks and rule based systems, in: Computational intelligence techniques in medical diagnosis and prognosis, SpringerLink, 2002.
[53] L. Peelen, N.F. de Keizer, E. Jonge, R.J. Bosman, A. Abu-Hanna, N. Peek, Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit, J. Biomed. Inform. 43 (2) (2010) 273–286.
[54] S. Curto, J.P. Carvalho, C. Salgado, S.M. Vieira, J.M.C. Sousa, Predicting ICU readmissions based on bedside medical text notes, in: Paper presented at: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); 24-29 July 2016, 2016.
[55] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, Crit. Care Med. 46 (4) (2018) 547–553.