Analysis of medical data using machine learning methods (MLHC)

Michal Alayev, Lian Galanti

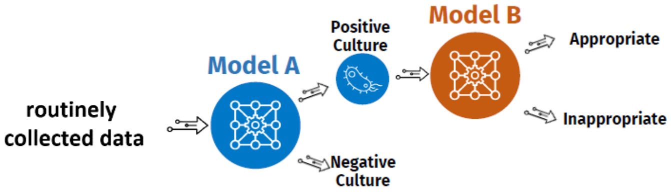
(1) Project Introduction

Bloodstream infections (BSI) account for significant morbidity and mortality in patients in the intensive care unit (ICU). Bacterial infection can be treated by antibiotics, but some bacteria are resistant to certain types. In case of viral or fungal infection, antibiotics are not useful and can even cause damage.

BSI is detected by taking a blood culture, the results for indication if the infection is bacterial or not are received after 24 hours, and the results for indication whether the infection is resistant or not to each type of tested antibiotic are received after 72 hours. In this period, the physicians need to set the treatment based on the patient's clinical presentation, in order to prevent the patient's deterioration. But when the blood culture results arrive, the treatment might be found as inappropriate, since the results may indicate that no antibiotics were needed, or that the antibiotic administered was wrong since the bacteria is resistant to it.

Since the ability of physicians to diagnose the infection characteristics when it is first observed is limited, our project objective was to develop two site-generalized ML-based classification models, using routinely collected data as input:

- **Model A -** classification model of positive vs. negative blood cultures of patients with ICU-acquired infections, based on clinical data that was collected before culture collection time.
- **Model B -** classification model of appropriate vs. inappropriate treatment that was given to patients with positive blood cultures, based on the clinical data that was collected until 24 hours after culture collection time.



(2) Cohort Description

Data sources

two data sets were used for developing the models**:**

MIMIC III - a dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data on ~60,000 intensive care unit admissions. It includes demographics, vital signs, laboratory tests, medications, and more.

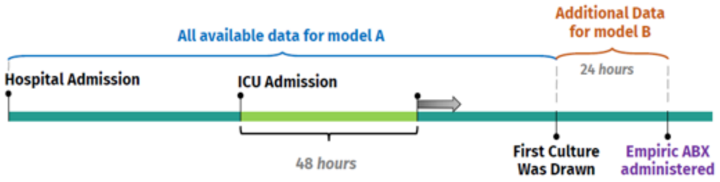
eICU - a dataset from many critical care units throughout the United States, comprising de-identified health data on ~200,000 patients who were admitted to critical care units in 2014 and 2015.

Inclusion Criteria

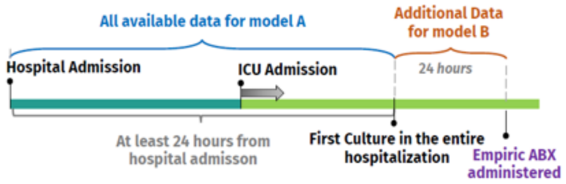
The cohort included patients whose blood culture was not contaminants or cancelled, that were admitted directly to Emergency department or ICU, and had ICU-acquired infection or suspected infection, as defined below:

- ICU-acquired Infection- patients who were hospitalized at least 48 hours in the ICU, and the first culture was collected in the ICU and after at least 48 hours from ICU admission. We considered only the first culture that was collected in the ICU.

- Suspected Infection- patient whose first culture in the entire hospitalization was collected in the ICU, and the culture was collected at least 24 hours after hospital admission.



Inclusion Criteria Timeline - ICU-acquired Infection


All available data for model A
Additional Data for model B
24 hours
Hospital Admission
ICU Admission
At least 24 hours from hospital admisson
First Culture in the entire hospitalization
Empiric ABX administered

Inclusion Criteria Timeline - Suspected Infection

Characteristics of the cohorts

Positive class is defined if there was a growth of pathogen in the culture of the patient.
In Model A class 1 represents the positive patients, and class 0 represents the negative patients.
In Model B class 1 represents patients that got inappropriate treatment, and class 0 represents those who got appropriate treatment.

	MIMIC Model A	MIMIC Model B	eICU
Number of patients	3075	105	126
Class 1 (% of patients)	161 (5.24)	22 (20.95)	8 (6.35)
Class 0 (% of patients)	2914 (94.76)	83 (79.05)	118 (93.65)
Age, years, median	66	69	69
Gender, male, %	1772 (57.63)	65 (61.90)	68 (53.97)
ICU stay length, median days	8.23	9.27	8.01
Hospital stay length, median days	15.59	18.37	15.19
Hospital mortality (% of patients)	683 (22.21)	31 (29.52)	34 (26.98)
Hospital mortality within class 1 (% of class 1)	52 (32.30)	12 (54.55)	3 (37.5)
Hospital mortality within class 0 (% of class 0)	631 (21.65)	19 (22.89)	31 (26.27)

Comparisons of demographic, clinical and laboratory variables at the time of blood culture sampling among positive vs. negative patients in MIMIC for model A:

MIMIC Model A		
Characteristics	**Negatives (n=2914)**	**Positives (n=161)**
Demographic and clinical status at admission		
Age (years), median (IQR)	66 (53, 78)	69 (55, 78)

Sex, male (%)	1673 (57.41)	99 (61.49)
ICU length of stay (days), median (IQR)	8.19 (4.71, 14.89)	9.21 (5.75, 16.45)
Hospital length of stay (days), median (IQR)	15.53 (9.58, 24.55)	16.13 (10.62, 27.11)
Clinical status during the 24 hours prior to BC sampling time		
There was an insertion of central venous line (%)	290 (9.95)	2 (1.24)
Got antibiotic treatment (%)	578 (19.84)	1 (0.62)
Were treated with vasopressors (%)	706 (24.23)	2 (1.24)
Vital signs, last values that were measured prior BC sampling time, median (IQR)		
Heart rate (bpm)	92.0 (80.0, 105.0)	91.0 (80.5, 104.5)
Respiratory rate (insp/min)	20.0 (16.0, 25.0)	19.5 (16.0, 23.75)
Arterial systolic blood pressure (mmHg)	121.0 (105, 142)	123.0 (109.5, 125.5)
Arterial diastolic blood pressure (mmHg)	61.0 (53.0, 71.0)	60.0 (53.5, 65.5)
Laboratory results, last values that were measured prior BC sampling time, median (IQR)		
Lymphocytes (%)	9.3 (5.7, 15.2)	10.0 (6.0, 17.0)
Neutrophils (%)	82.95 (74.73, 88.48)	81.9 (71.25, 87.25)
pH	7.41 (7.37, 7.45)	7.39 (7.34, 7.43)
Lactate (mmol/L)	1.5 (1.1, 2.1)	1.55 (1.1, 2.58)
pO2 (mmHg)	106.0 (82.0, 140.0)	98.5 (80.0, 136.5)
Albumin (g/dL)	3.0 (2.5, 3.5)	2.8 (2.5, 3.4)

Comparisons of demographic, clinical and laboratory variables at the time of blood culture sampling among positive vs. negative patients in eICU for model A:

eICU Model A		
Characteristics	**Negatives (n=118)**	**Positives (n=8)**
Demographic and clinical status at admission		
Age (years), median (IQR)	69 (57, 77)	63 (47, 73)
Sex, male (%)	64 (54.24)	4 (50.0)
ICU length of stay (days), median (IQR)	8.01 (5.39, 13.63)	7.01 (5.08, 12.29)
Hospital length of stay (days), median (IQR)	14.72 (9.81, 19.99)	21.74 (12.42, 27.53)
Clinical status during the 24 hours prior to BC sampling time		
There was an insertion of central venous line (%)	8 (6.78)	0 (0.0)
Got antibiotic treatment (%)	21 (17.80)	4 (50.0)
Were treated with vasopressors (%)	35 (29.66)	3 (37.5)
Vital signs, last values that were measured prior BC sampling time, median (IQR)		
Heart rate (bpm)	92.0 (78.0, 105.5)	95.0 (88.25, 101.75)
Respiratory rate (insp/min)	21.0 (17.5, 25.0)	21.0 (20.0, 22.5)
Arterial systolic blood pressure (mmHg)	123.0 (113.5, 142.25)	103.0 (101.0, 105.0)
Arterial diastolic blood pressure (mmHg)	58.0 (51.0, 70.75)	39.0 (33.5, 44.5)

Laboratory results, last values that were measured prior BC sampling time, median (IQR)		
Lymphocytes (%)	8.8 (4.0, 14.53)	8.4 (5.65, 14.1)
Neutrophils (%)	80.0 (72.0, 86.25)	78.6 (67.8, 86.8)
pH	7.41 (7.34, 7.47)	7.38 (7.37, 7.47)
Lactate (mmol/L)	1.5 (1.1, 2.3)	1.5 (1.3, 1.6)
pO2 (mmHg)	86.0 (67.0, 112.5)	136.0 (83.0, 175.5)
Albumin (g/dL)	2.6 (2.1, 3.1)	2.1 (1.88, 2.48)

Comparisons of demographic, clinical and laboratory variables at the time of blood culture sampling among appropriate vs. inappropriate patients in MIMIC for model B:

MIMIC Model B		
Characteristics	**Appropriate (n=83)**	**Inappropriate (n=22)**
Demographic and clinical status at admission		
Age (years), median (IQR)	68 (55, 77)	69 (60, 78)
Sex, male (%)	54 (65.06)	11 (50.0)
ICU length of stay (days), median (IQR)	9.96 (5.92, 17.29)	7.80 (4.48, 18.17)
Hospital length of stay (days), median (IQR)	18.4 (10.05, 26.55)	15.16 (7.76, 24.98)
Clinical status during the 24 hours prior to BC sampling time		
There was an insertion of central venous line (%)	2 (2.41)	2 (9.09)
Got antibiotic treatment (%)	5 (6.02)	3 (13.64)
Were treated with vasopressors (%)	11 (13.25)	3 (13.64)
Vital signs, last values that were measured prior BC sampling time, median (IQR)		
Heart rate (bpm)	89.0 (78.5, 102.0)	95.0 (77.5, 105.25)
Respiratory rate (insp/min)	21.0 (16.0, 25.0)	23.0 (18.25, 27.5)
Tidal Volume (spontaneous) (mL)	440.0 (361.0, 561.5)	529.5 (405.5, 606.0)
Arterial systolic blood pressure (mmHg)	117.5 (98.55, 131.5)	106.0 (104.0, 108.0)
Arterial diastolic blood pressure (mmHg)	57.0 (49.25, 69.0)	52.0 (49.0, 55.0)
Laboratory results, last values that were measured prior BC sampling time, median (IQR)		
Lymphocytes (%)	6.95 (4.0, 12.43)	5.0 (2.7, 13.3)
Neutrophils (%)	83.0 (77.0, 88.5)	81.0 (76.05, 91.20)
pH	7.42 (7.38, 7.45)	7.42 (7.29, 7.44)
Lactate (mmol/L)	1.4 (1.2, 2.15)	1.70 (1.20, 3.15)
pO2 (mmHg)	108.0 (89.5, 126.5)	112.0 (79.0, 137.0)
Albumin (g/dL)	3.0 (2.5, 3.53)	2.45 (1.83, 3.2)

Missing Data

In all three datasets (MIMIC for model A, MIMIC for model B, and eICU for model A) there was a significant lack of data regarding antibiotics treatment, pressor-sedatives treatment, presence of invasive lines, and liquid input (above 70% of the patient in each cohort did not have any records regarding the desired data). Since we thought this information might be of great importance for the prediction, we generated binary features, where 0 was the default for patients with no data regarding the created feature.

For most laboratory and respiratory parameters, 0%-30% of the patient in each cohort did not have any records regarding the desired data, but for vital signs such as arterial blood pressure and temperature the missing rates were high (77% and 76% respectively in MIMIC model A). Elaboration of the missing data rates in the different datasets as the percentage of cohort patients without any documentation of the variable during the hospitalization prior blood culture sampling, is presented in the supplemental material 1 (SM1).

(3) Methods

Data Preprocessing

a. Data loading: for both models, we created datasets that included demographics, clinical and laboratory parameters, GCS scores, microbiology tests, medical treatment that was provided (pharmacologic treatment and presence of indwelling catheters), and daily fluid input and output, from the data that was available in the databases between admission and BC sampling time.

The eICU dataset was meant to be used as validation data for model A that was trained on MIMIC dataset, hence the generated features for model A were limited to the data that was available in both databases, to ensure compatibility.

In MIMIC database, the data of prior medical diagnoses, surgeries during the admission, prior comorbidities, and prior treatment with immunosuppressive drugs was unavailable. In eICU database, data of ventilation events was rare for the patients in the cohort. Due to these limitations, those features were not generated.

Model B includes overall similar data to model A.

The parameters included in the datasets for both models are generally elaborated in SM 2.

b. Data cleaning: we defined plausible ranges for vital signs, laboratory tests, and respiratory parameters. Implausible values were excluded, as it is plausible that this data is imprecise and using it could harm the model.

Description of the defined plausible ranges is shown in SM 3.

c. Feature extraction: the data includes many parameters that change over time, such as lab measurements, vital signs and respiratory parameters. Each feature was preprocessed to represent these changes between hospital admission and the BC sampling time, by generating different statistics of the values (the last, maximum, minimum, average, median, 25th and 75th percentiles, standard deviation, number of measurements, etc.), and generating features that represented the time (hours) that has elapsed between the BC sampling time and the time that the statistical values (maximum, minimum, etc.) were recorded.

Similar processing and feature generation was done for GCS scores, weight (when measured daily), daily liquid input and output amounts, since they also change over time.
For pharmacological treatment, presence of catheters and lines, and microbiology tests, we generated mostly binary features, with indicate whether a patient got a specific treatment or procedure (line insertion, collection of a culture, etc.)

Modeling

For both models, the created datasets consisted of over 1000 generated features.
MIMIC datasets were used for training the models, and eICU dataset was used as validation set for model A.

Model A:

We first removed features that had at least 70% missing values rate.
Next, we split the MIMIC dataset into 5 fixed groups for 5-fold cross-validation, so in each fold about 80% of the dataset was used as train data, and the other 20% was used as test data.
The train and test data were imputed separately with KNN using 10 nearest neighbors and uniform weight function. To handle the class imbalance issue, The train data was balanced using Borderline SMOTE for oversampling (adding positive samples).
The train and test data were then separately standardized by removing the mean of each feature and scaling to unit variance, using Standard Scaler.
Following these processing steps, we applied feature selection to the train data, keeping the 40 features with highest scores according to ANOVA F-value, to avoid overfitting (model recognition of specific cases instead of generalizing).
For classification, we used logistic regression with the lbfgs solver and norm 2 for penalization.

Model B:

We first removed features that had at least 70% missing values rate, as in model A. Then we removed features with variance lower than 0.15.
The dataset with the remaining features was split into 5 fixed groups for 5-fold cross-validation.
The train and test data in each run were imputed separately with KNN using 5 nearest neighbors and uniform weight function. The train data was balanced using Tomek Links for undersampling (removing appropriate treatment samples), in order to handle the class imbalance. The train and test data were separately standardized by removing the mean of each feature and scaling to unit variance, using Standard Scaler.
Following these processing steps, we applied a logistic regression feature selection algorithm on the train data, with the lbfgs solver and norm 2 for penalization. We kept the 10 features with the greatest coefficient (negative or positive).
For classification, we used logistic regression with the lbfgs solver and norm 2 for penalization, as in model A.

Evaluation Process

We used 5-fold cross-validation method on the MIMIC datasets of each model. Each dataset was divided into 5 parts according to a random predefined partition, and we let the model run five times. In every such run, the model uses 4 different parts of the training set for modeling and the fifth part for validation.

In addition, we made a second evaluation for model A, when the train data was the MIMIC dataset, and the test data was the eICU dataset. This required both datasets to have the same features, as we designed in the feature generation step.

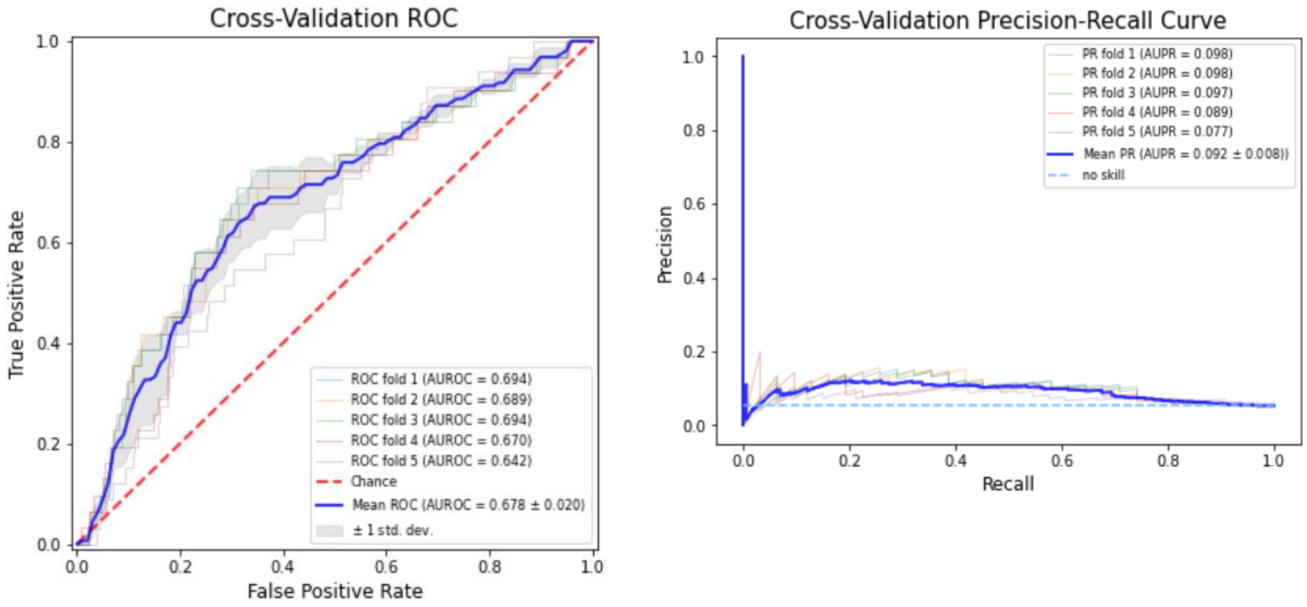
(4) **Results**

Classification performance was assessed through the computation of the area under the receiver operating characteristics (AUROC) and the area under the precision-recall curve (AUPR) for each fold, and then calculating the mean AUROC and AUPR.

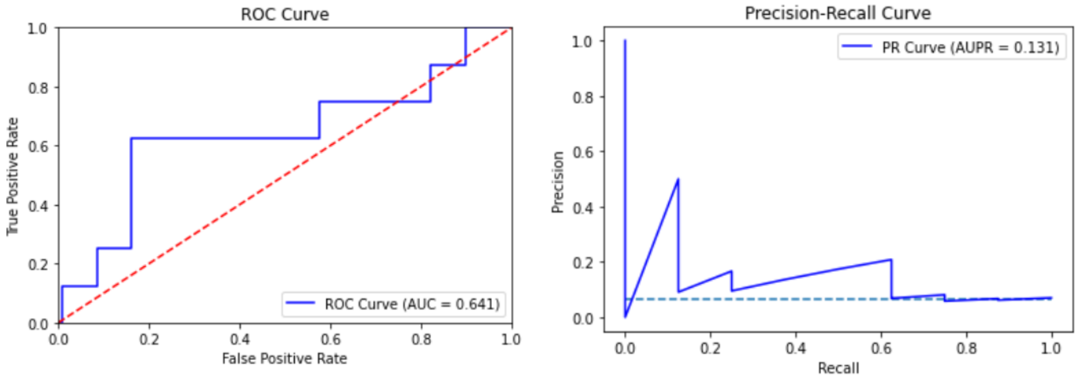
On model A, the cross-validation AUROCs were 0.678 ± 0.020 and AUPRs were 0.092 ± 0.008, while the validation set (eICU dataset) AUROC was 0.641 and the AUPR was 0.131.

On model B, the cross-validation AUROCs were 0.759 ± 0.050 and AUPRs were 0.315 ± 0.001.

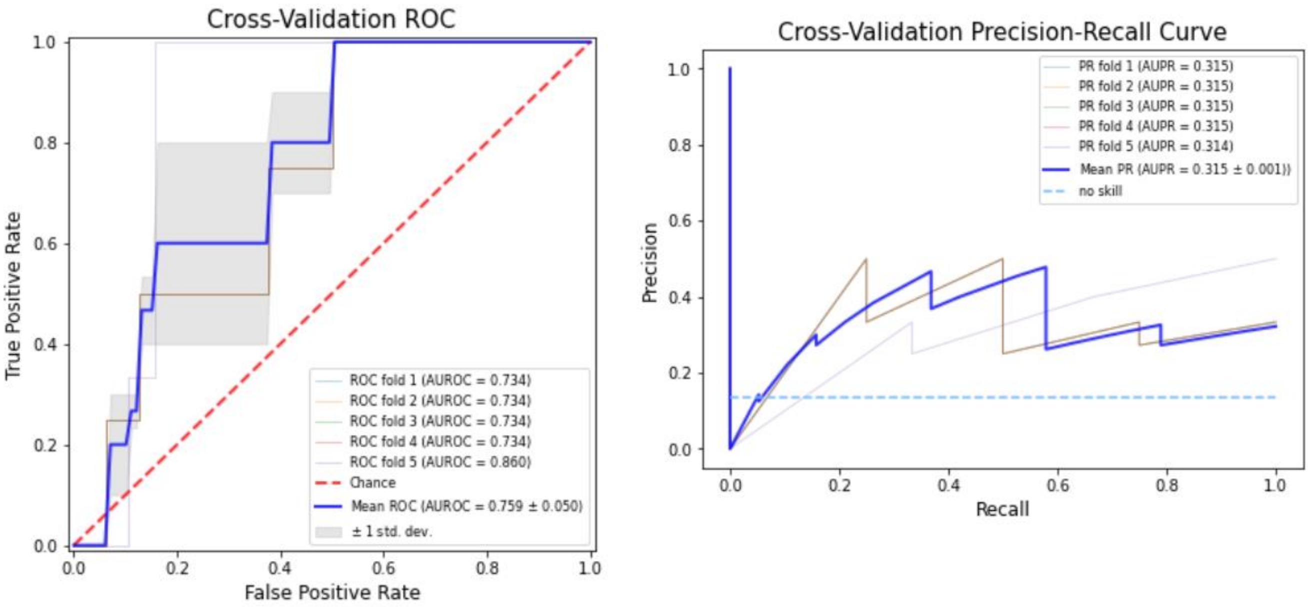
Cross-Validation results for model A (MIMIC dataset)



External validation set results (on eICU dataset) for model A



Cross-Validation results for model B (MIMIC dataset)



In model A, the 40 selected features most predictive of positive blood cultures included mainly features of antibiotic and pressor-sedatives treatment, presence of arterial and central venous lines, and some features that represent statistics of laboratory and vital signs measurements.
In model B, the 10 selected features most predictive of inappropriate treatment included mainly features of respiratory data.
All of the selected features in each model are elaborated in SM 4, with F-value and p-value for model A, and logistic regression coefficients for model B.

(5) **Discussion**

In this project, we developed ML algorithms to predict bacterial BSI among patients with ICU-acquired or suspected bloodstream infection, and to predict appropriate vs. in appropriate treatment among patients with positive blood culture. The algorithms achieved good prediction in cross-validation, and in external validation for model A. The goal was to develop a robust algorithm that can be used for different medical databases, rather than site-specific algorithm. The results of model A on eICU dataset show that this goal is achievable, even though the two different datasets represent different medical centers with different patient characteristics, bacterial BSI incidence, and different database structure.

Our project has several limitations. Our prediction modeling in model A starts with patients suspected of bacterial BSI, for whom blood cultures were taken. The decision of the attending physician to take blood cultures was based on clinical judgment. In some cases, bacterial BSI may have been overlooked, while superfluous testing may have been performed in other cases.

In addition, we excluded all episodes of blood culture collection with the growth of potential contaminants. It is possible that some of those were true cases of bacterial BSI.
Another limitation was in model B- since it includes only patients with positive blood culture, the cohort size for model was very small, and due to the very few positive patients in eICU database (only 8 patients), model B could not be tested on eICU as external validation set. This limited the inspection of model B as a generalized prediction model, rather than site-specific.

In conclusion, we developed algorithms that uses easily accessed, local electronic clinical data of patients, to identify patients who are at high risk for bacterial BSI among suspected ICU-acquired infection, and to identify patients that are most likely to be treated inappropriately among patients with bacterial BSI. With further improvements and research, the models' performance could be high enough to serve as a support tool for the decisions on whether to start or to what extent broaden/change antibiotic treatment.

(6) **References**

1 Roimi, M., Neuberger, A., Shrot, A. *et al.* Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms . *Intensive Care Med* **46,** 454–462 (2020). https://doi.org/10.1007/s00134-019-05876-8