BMJ Open

Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study

Anneroos W Boerman,[1,2] Michiel Schinkel,[1,3] Lotta Meijerink,[4] Eva S van den Ende,[1] Lara CA Pladet,[1] Martijn G Scholtemeijer,[4] Joost Zeeuw,[4] Anuschka Y van der Zaag,[1] Tanca C Minderhoud,[1] Paul W G Elbers,[5] W Joost Wiersinga,[3,6] Robert de Jonge,[2] Mark HH Kramer,[7] Prabath W B Nanayakkara [1]

To cite: Boerman AW, Schinkel M, Meijerink L, *et al*. Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study. *BMJ Open* 2022;**12**:e053332. doi:10.1136/bmjopen-2021-053332

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-053332).

AWB and MS contributed equally.

Received 13 May 2021
Accepted 03 December 2021

For numbered affiliations see end of article.

Correspondence to
Professor Prabath W B Nanayakkara;
p.nanayakkara@amsterdamumc.nl

ABSTRACT

Objectives To develop predictive models for blood culture (BC) outcomes in an emergency department (ED) setting.
Design Retrospective observational study.
Setting ED of a large teaching hospital in the Netherlands between 1 September 2018 and 24 June 2020.
Participants Adult patients from whom BCs were collected in the ED. Data of demographic information, vital signs, administered medications in the ED and laboratory and radiology results were extracted from the electronic health record, if available at the end of the ED visits.
Main outcome measures The primary outcome was the performance of two models (logistic regression and gradient boosted trees) to predict bacteraemia in ED patients, defined as at least one true positive BC collected at the ED.
Results In 4885 out of 51 399 ED visits (9.5%), BCs were collected. In 598/4885 (12.2%) visits, at least one of the BCs was true positive. Both a gradient boosted tree model and a logistic regression model showed good performance in predicting BC results with area under curve of the receiver operating characteristics of 0.77 (95% CI 0.73 to 0.82) and 0.78 (95% CI 0.73 to 0.82) in the test sets, respectively. In the gradient boosted tree model, the optimal threshold would predict 69% of BCs in the test set to be negative, with a negative predictive value of over 94%.
Conclusions Both models can accurately identify patients with low risk of bacteraemia at the ED in this single-centre setting and may be useful to reduce unnecessary BCs and associated healthcare costs. Further studies are necessary for validation and to investigate the potential clinical benefits and possible risks after implementation.

Strengths and limitations of this study

- These models are based on routinely collected clinical data that are available at the end of a visit at the emergency department and are therefore applicable to implement in clinical practice.
- Free-text data, such as physician and nurse reports, could not be used due to privacy concerns.
- These models should not be used in patients at high risk for bloodstream infections caused by pathogens that are usually reported as contaminants, such as with central line associated infections.
- This is a single-centre study and further studies on validation and implementation are necessary to investigate possible risks and likely benefits of these models.

INTRODUCTION

Over 20% of adult emergency department (ED) visits occur due to serious infections.[1] Current diagnostic modalities cannot sufficiently distinguish between bacterial and non-bacterial disease during an early stage of a diagnostic workup, for instance in case of a possible bacteraemia (bloodstream infection).[2] However, timely distinction between bacterial and non-bacterial disease can reduce unnecessary diagnostic tests and treatment with antibiotics. In case of a bacteraemia, blood cultures (BCs) are the gold-standard test. Unfortunately, turnaround times of BC results of 24–72 hours make these cultures unhelpful for timely diagnosis of bacterial infections at the ED. Accurate and early identification of patients with a high or low risk of bacteraemia may be a first step to help distinguish bacterial from non-bacterial disease early.

Bacteraemia is associated with high morbidity and mortality, which makes missing a possible bacteraemia very harmful.[3] Therefore, physicians order BCs frequently and the overall BC yields are low.[2] Around 11%–15% of collected BCs are positive and studies show that up to half of those are false positives through contamination.[4–6] These contaminated BCs can also lead to unnecessary downstream diagnostics, antibiotic overuse and increased hospital length of stay.[7–9] Currently, we are unable to recognise patients with low risk of bacteraemia, in which we could safely withhold BC testing and even antibiotics.

Machine learning already has significant impact on healthcare. Machine learning models can use many data points from large numbers of patients to detect subtle patterns that may go unnoticed by healthcare professionals. These insights may support the swift assessment of a patient and selection of the appropriate diagnostic and treatment strategies. Complex situations, where multiple physiological mechanisms interact are perfect areas to investigate machine learning decision support.[10] The diagnostic workup of suspected bacterial infections is such an area.

In this paper, we aim to create predictive models for BC outcomes in the ED setting which may help reduce unnecessary BCs and provide physicians with an additional tool to help decide whether or not antibiotic treatment is needed. We specifically focus on creating a machine learning pipeline that can be easily adapted and available in many settings.

METHODS

Study setting

We performed a retrospective observational study on data from the electronic health records (EHR) of Amsterdam UMC, location VU University Medical Center, between 1 September 2018 and 24 June 2020. The VU University Medical Center is a large teaching hospital with an estimated 28 000 ED presentations annually. The study adhered to the ‘transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)’.[11]

Population

We included all adult patients who presented to the ED and in whom at least one BC was taken during their ED stay because a bacterial infection was suspected on clinical grounds. We included patients of all medical specialties. Whenever a patient presented to the ED multiple times during the study period, each encounter was classified as a unique visit.

Data collection

All data that were available under local privacy regulations was extracted from the EHR. The data included demographic information, vital signs, laboratory results and information about imaging procedures and administered medications in the ED. Data on comorbidities or medication usage at home were not available. We only used data that would be available before the end of the ED visit, which is the time when the prediction can potentially have clinical consequences on the use of BCs and initiation of antibiotic therapy. The data extracted from the EHR was further preprocessed to be used for predictive modelling. Details about preprocessing are described in online supplemental e-methods and e-tables 1–4.

Outcome

We aimed to predict bacteraemia, which was defined as at least one positive BC with a pathogenic microorganism collected during the ED visit.

AWB and MS mapped all microorganisms to be pathogens or contaminants based on previous literature under supervision of WJW.[2 4 12 13] Online supplemental e-table 5 lists all organisms that we classified as contaminants. Then, we assigned the most important result to a specific BC set (prioritising positive over contamination over negative). Afterwards, the combination of all BC sets in a unique ED visit was mapped to represent a visit with growth of a clinically significant pathogen in at least one BC set (positive) or a visit with only negative or contaminated cultures (negative).

Model development and feature selection

We used all variables that were reported in over 10% of the ED visits as features. We also created indicator features for all variables to indicate whether this variable was measured or not. The dataset was randomly split into a training (75%) and test (25%) set for model development. We used median imputation except for some situations where imputation based on domain knowledge was used (see online supplemental etable 6 for details). Median imputation is a practical and adequate solution for handling missing data in non-linear models. Furthermore, the combination of median imputation and indicator features as we used is also adequate for linear models, especially with data missing not at random.[14] Additional standard scaling around the mean was applied. We trained the models on the training set using the full set of features, since the used models are robust to unimportant features.

We used a gradient boosted tree model and a logistic regression model with L1 regularisation. These different model classes are known to be suitable for our type of data, which is limited in size and of mixed type. We used gradient boosted trees as a powerful representative of tree-based models, which can uncover complex feature interdependencies and non-linearities. We also used a simpler logistic regression for comparison, since its coefficients are easier to interpret.

Within the training set, a fivefold cross-validated grid search was performed to find the hyperparameters that optimise the model’s performances. An overview of the pipeline from raw data to model can be found in the e-methods section of online supplemental appendix.

Modelling was performed using Python V.3.7.9 (Python software foundation, http://www.python.org) and the Scikit-learn package (V.23.1).

Model evaluation

The model performances were tested using the area under curve of the receiver operating characteristics (AUROC), together with the area under the precision recall curve (AUPRC) since we had imbalanced outcome classes. We also reported Brier scores and F1-scores during cross-validation as well as on the test set. The model calibration is presented in calibration plots.

The model's output was the probability for the BC to be positive. To provide a clinically meaningful result, we report on two preselected probability thresholds that predict BCs to be positive above this threshold. First, we show performances on the most optimal sensitivity-specificity threshold based on maximisation of the sensitivity-specificity sum or minimisation of the sensitivity-specificity difference.[15] These approaches are useful when omission errors (false negatives) should be avoided and provide a diagnostic test with the power to rule out a diagnosis.[15 16] Furthermore, we present model performances on a threshold that retains a sensitivity of 90%, which is in line with our goal of using it to identify patients in which we can safely withhold collecting a BC.

Patient and public involvement

Patients were not involved in setting the research question, design of the study, outcome measures and interpretation of the study.

RESULTS

Baseline characteristics

We identified 51 399 ED visits by 41 280 unique adult patients in the VU University Medical Center between 1 September 2018 and 24 June 2020. One or more BC samples were taken in 4885 (9.5%) of those visits. In 598/4885 (12.2%) of those visits, at least one of the cultures was a true positive. In 254/4885 (5.2%) of the visits, at least one of the cultures was contaminated (later mapped to be negative). Overall, 4074/4885 (83.4%) visits had only truly negative cultures. Table 1 shows the baseline characteristics of the study population stratified by culture outcomes.

Predictive performance

The gradient boosted tree model's AUROC in the cross-validation (training) sets and internal test set were 0.77 (SD=0.03) and 0.77 (95% CI 0.73 to 0.82), respectively. The logistic regression model's AUROC in the cross-validation and internal test set were 0.75 (SD=0.02) and 0.78 (95% CI 0.73 to 0.82). The AUROCs of both models are shown in figure 1. Table 2 shows the corresponding performance scores. The calibration plots are presented in online supplemental e-figure 1.

Feature importances

Gradient boosted trees

Feature importances for non-linear tree based models only indicate the magnitude and not the directionality (positive/negative) of the effect. We present the feature contributions using shapley additive explanation values, as depicted in figure 2.[17] These are distributions of local contributions per feature and per data point. Figure 2 shows the 20 most important features that drive predictions in the gradient boosted tree model (see online supplemental e-table 2–4 for the full lists of features). This model recognises bilirubin values to be the strongest predictor of a positive BC. We see that high (red) bilirubin values are associated with a higher risk of a positive BC (right on the x-axis). Conversely, high (red) potassium levels are associated with a lower risk of a positive BC (left on the x-axis).

Logistic regression

The 20 features with the largest absolute coefficients in the logistic regression model are presented in figure 3. Age and lymphocyte counts are the strongest predictors. A high age is associated with a higher risk of a positive BC, whereas a high lymphocyte count is associated with a lower risk (see online supplemental e-table 7 for a full list of coefficients). Due to the imputation and the fact that physiological parameters are not strictly independent of each other, no valid estimation of the ORs can be provided.

Thresholds

The models sensitivity and specificity depend on the probability threshold that is used to predict a positive or negative BC. Table 3 presents model performances for the optimal sensitivity-specificity threshold and a threshold that retains a sensitivity of 90%. The optimal threshold in the gradient boosted tree model would predict 69% of BCs in the test set to be negative, with a negative predictive value of over 94%. An extensive list of thresholds and corresponding performances in both sets can be found in online supplemental e-table 8 and 9.

Medication administered in the ED

In coming to the final models, we evaluated the effects of excluding different groups of features, such as medications given in the ED. Excluding all ED medication features led to comparable model performances (see online supplemental e-table 10 for details). When including the ED medication features, almost none provided predictive value, except for the administration of antibiotics (see online supplemental e-figures 2 and 3). Because this event may be associated with the physician's suspicion of bacteraemia, we decided to exclude ED medication features in order to retain a model that can augment physician decision making instead of depending on it.

DISCUSSION

We present two models that aim to predict the outcome of a BC that is drawn during an ED visit. Both a gradient boosted tree model and a logistic regression model show comparably good performance in predicting BC results with AUROCs of 0.77 (95% CI 0.73 to 0.82) and 0.78 (95% CI 0.73 to 0.82) in the test sets, respectively. In a

Table 1 Baseline characteristics of the study population stratified on blood culture outcomes

Characteristic	Negative cultures* (N=4287)	Positive cultures (N=598)	Total (N=4885)
Age, years			
Median (IQR)	66 (51–75)	70 (59–79)	66 (52–76)
Sex			
Male	56.3%	62.2%	57.0%
Modified Early Warning Score			
Median (IQR)	3 (2–4)	4 (2–5)	3 (2–4)
Missing (N)	2515	351	2866
Heart rate (beats per minute)			
Median (IQR)	94 (82–107)	100 (88–111)	95 (82–108)
Missing (N)	181	15	196
Systolic blood pressure (mm Hg)			
Median (IQR)	126 (112–142)	118 (104–136)	125 (111–141)
Missing (N)	372	36	408
Respiratory rate (per minute)			
Median (IQR)	19 (15–23)	21 (16–25)	19 (15–24)
Missing (N)	1310	149	1459
Temperature (degree celsius)			
Median (IQR)	37.8 (37.0–38.5)	38.1 (37.2–38.8)	37.8 (37.0–38.5)
Missing (N)	198	26	224
C reactive protein (µmol/L)			
Median (IQR)	60 (25–134)	104 (39–216)	64 (25–144)
Missing (N)	132	23	155
Whitecell counts (10^9/L)			
Median (IQR)	10 (6.8–13.8)	11.9 (8.2–16.0)	10.2 (6.9–14.2)
Missing (N)	144	22	166
Thrombocyte counts (10^9/L)			
Median (IQR)	234 (174–311)	211 (149–273)	231 (171–307)
Missing (N)	593	105	698
Bilirubin (µmol/L)			
Median (IQR)	9 (6–13)	13 (8–22)	9 (6–14)
Missing (N)	1205	163	1368
Creatinine (µmol/L)			
Median (IQR)	82 (65–113)	105 (73–160)	84 (66–119)
Missing (N)	171	27	198
Length of ED stay (hours)			
Median (IQR)	4.3 (3.2–5.8)	4.7 (3.3–6.3)	4.4 (3.2–5.9)
Hospital admission			
Admitted	68.0%	84.6%	70.0%
30-day mortality			
Died	6.7%	11.5%	7.3%

*Likely contaminants are classified as negative cultures in this table.
ED, emergency department.

population where the physicians has made the decision to draw a BC, the models can identify patients in the ED with low risk for bacteraemia and can be useful to reduce unnecessary BCs and provide physician decision support on the necessity of antibiotic therapy.

Many studies have aimed to identify factors associated with positive BCs or predict BC outcomes. A 2012 systematic review reported on 35 studies that evaluated the performance of clinical variables to detect bacteraemia.[2] Those clinical variables alone seemed insufficient

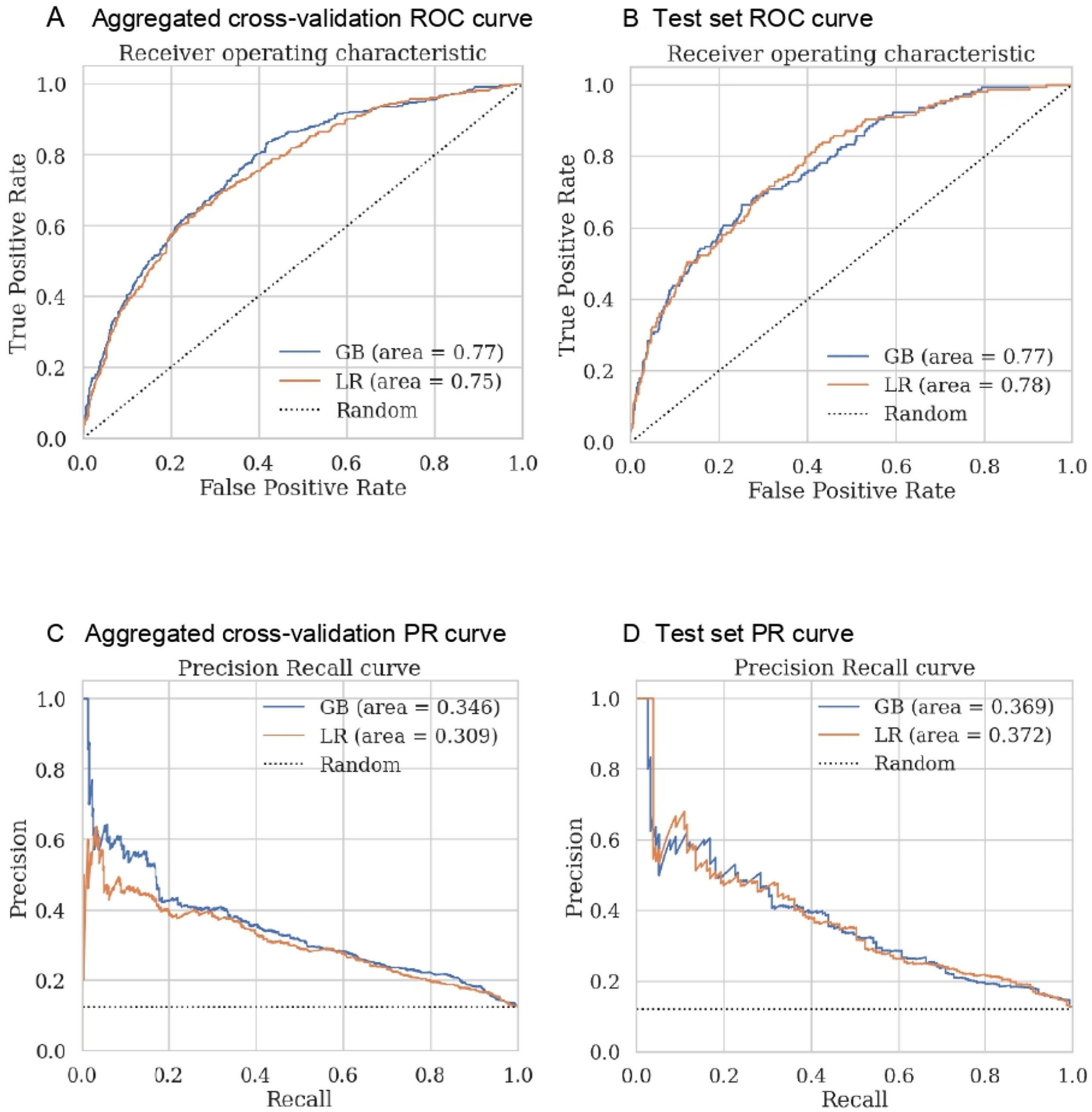

Figure 1 Receiver operating characteristic (ROC) and precision recall (PR) curves for positive blood cultures in aggregated cross-validation sets and test set. GB, gradient boosted tree model; LR, logistic regression model.

to detect bacteraemia and further studies on this subject have focused on more advanced predictive models to detect bacteraemia. A 2015 systematic review presented fifteen machine learning models that predicted BC outcomes.[18] An additional few were published since.[19–22]

The various studies on this subject have been conducted in different settings, where the reasons for drawing BCs vary. We focused on the ED setting, as the legacy of a probable diagnosis of infection at the ED greatly influences decision-making throughout the hospital stay,

Table 2 Performance metrics of both models in the aggregated cross-validation sets and the test set

Model	Modelling phase	AUROC	AUPRC	Brier score*	F1-score†
Gradient boosted trees	Cross-validation mean	0.77 (SD=0.03)	0.340	0.066	0.16
	Test	0.77 (95% CI 0.73 to 0.82)	0.37	0.092	0.17
Logistic regression	Cross-validation mean	0.75 (SD=0.02)	0.31	0.098	0.14
	Test	0.78 (95% CI 0.73 to 0.82)	0.37	0.092	0.16

*The Brier score is a cost function that measures performance of probabilistic predictions. The score ranges from 0 to 1. The lower the score, the more accurate the prediction.
†F1-scores present a balance between precision and recall. The higher the score, the more accurate the prediction.
AUPRC, area under the precision recall curve; AUROC, area under the curve of the receiver operating characteristics.

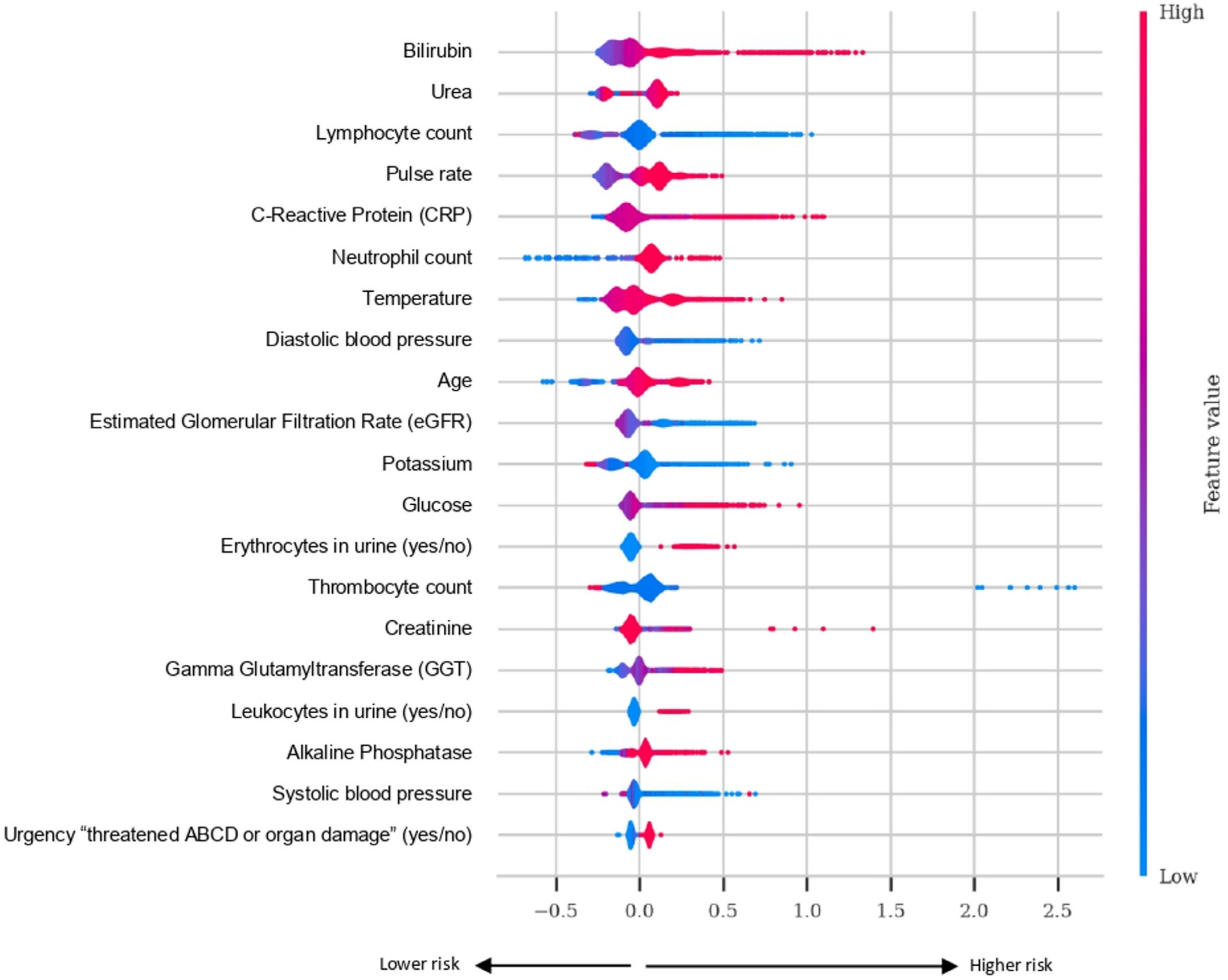

Figure 2 SHAP-plot of feature importance in the gradient boosted tree model. SHAP, shapley additive explanation. The figure shows the 20 most important features in the gradient boosted tree model, in order of importance on the Y-axis. The relative effect of the feature on the risk of a positive blood culture is shown on the X-axis (right of 0.0 = increased risk, left of 0.0 = lower risk). The colours represent the actual values of the features themselves. Blue depicts a low actual value of the feature while red depicts a high actual value. With yes/no features, no is depicted as a low value (blue) and yes as a high value (red).

especially with regards to antibiotic treatment.[23] Based on the 2015 systematic review, only two other studies have been carried out fully in an ED setting.[18 24 25] Those models showed AUROCs of 0.75 and 0.74 in the test sets. Of those two studies, the one by Shapiro *et al*[24] has had the most influence on clinical practice, as the Shapiro decision rule has been studied and used in hospitals around the world.[5 24 26] Our algorithms, with AUROCs of 0.77 and 0.78, perform at least as well as the Shapiro model and are only based on regularly captured EHR data. The major difference between previous models and our study is that those earlier models were trained on data that were prospectively collected by researchers. This manual data collection resulted in few missing values, with 97.6% of laboratory data being available.[24] This will not occur in clinical practice and may lead to dramatic losses in predictive performance in implementation studies, when missing values need to be imputed in order to do any prediction. Therefore, these models have less potential for daily use in clinical practice and it will be difficult to implement them successfully.

Another aspect of the manual data collection in earlier studies is that predictors like the suspicion of endocarditis, which was an important predictor of BC outcomes, could be used.[24] This is very specific data that will rarely be available in the EHR, which again limits the translation to clinical practice and automation of the prediction within an EHR environment. As we illustrate here, the use of data that is not routinely captured in clinical practice is one of the key reasons why none of these prediction models have been implemented in clinical practice yet.[18] In contrast, the overarching approach we used, with a machine learning pipeline that incorporates variables measured in certain percentages of

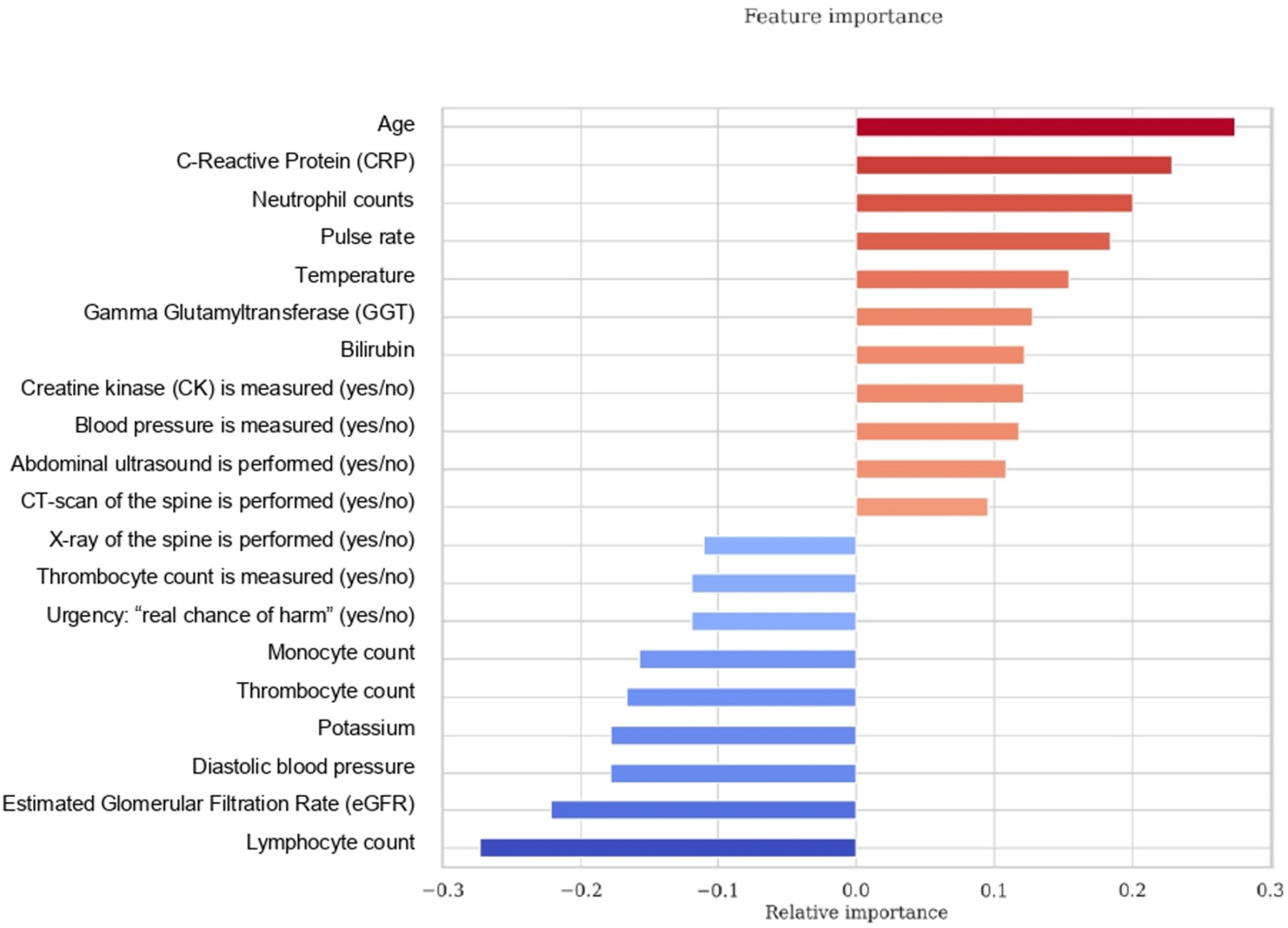

Figure 3 Feature importances of the logistic regression model. The 20 most important features in the logistic regression model are shown. The features for which a high value is predictive of a positive BC are shown in red and those predictive of a negative culture in blue. The X-axis presents the relative importance of these features.

patients, ensures that the use of this pipeline in other hospitals will produce usable models that are slightly adapted to the particular setting of that hospital. Our approach can thus straightforwardly be implemented in various setting in clinical practice, without the need of additional data capture.

Most of the literature on BC predictions focuses on the intensive care unit (ICU) setting. Recent examples are models created by Roimi *et al* and van Steenkiste *et al*.[20 21] Those models show excellent performances with AUROCs of up to 0.98 in the critical care setting. These models are trained on temporal trends that have occurred over a period of at least 48 hours, in contrast with the short and heterogeneous ED visits during which patients are not constantly monitored and where time-series data is rarely captured. Also, the approaches as taken for most intensive care unit (ICU) models seem to be overfitting to the training data and will likely perform worse in an external validation. This is underscored in the model by Roimi *et al*, in which the AUROC decreases from 0.92 to 0.60 during external validation.[21]

Clinical value

The main clinical value of our predictive model lies in the ability to identify patients at low risk of a positive BC, in a population where the physician has decided that a BC draw should be performed. The prediction can be made at the end of the ED visit and can identify patients in which we can safely withhold BC testing. Even in cases where BCs are already taken, there would be the option to not go through with the analyses, where most of the costs and associated harms are made. We showed that we would be able to withhold BC draws or analyses in almost 70% of the population while still retaining a negative predictive value of over 94%.

Our algorithm also has added value with regards to treatment selection, especially in cases with high diagnostic uncertainty at the end of an ED visit. The BC outcome prediction can be used as decision support tool to decide whether or not antibiotic treatment is needed. Estimated rates of unnecessary antibiotic use at the ED are over 30%, and it has been described as the most preventable cause of antibiotic resistance.[27] Predictions of negative BCs can be an additional argument for withholding

Table 3 Performance metrics for both models at preselected thresholds in the aggregated cross-validation sets and the test set

Model and metric	Optimal sensitivity-specificity Cross-validation (n=3608)	 Test (n=1277)	Sensitivity retained at over 90% Cross-validation (n=3608)	 Test (n=1277)
Gradient boosted tree model				
Threshold for positive prediction	10%	12.5%	6%	6%
True negative (n (%))	2126 (58.9)	829 (64.9)	1369 (37.9)	473 (37.0)
True positive (n (%))	322 (8.9)	103 (8.1)	400 (11.1)	142 (1.1)
False negative (n (%))	121 (3.4)	52 (4.1)	43 (1.2)	13 (1.0)
False positive (n (%))	1039 (28.8)	293 (22.9)	1796 (49.8)	649 (50.8%)
Sensitivity (%)	72.7	66.5	90.3	91.6
Specificity (%)	67.2	73.9	43.3	42.2
Positive predictive value (%)	23.7	26	18.2	18
Negative predictive value (%)	94.6	94.1	97	97.3
Logistic regression model				
Threshold for positive prediction	12.5%	10%*	6%	6%
True negative (n (%))	2172 (60.2)	680 (53.2)	1144 (31.7)	429 (33.6)
True positive (n (%))	308 (8.5)	123 (9.6)	405 (11.2)	142 (11.1)
False negative (n (%))	135 (3.7)	32 (2.5)	38 (1.1)	13 (1.0)
False positive (n (%))	993 (27.5)	442 (34.6)	2021 (56.0)	693 (45.3)
Sensitivity (%)	69.5	79.4	91.4	91.6
Specificity (%)	68.6	60.6	36.1	38.2
Positive predictive value (%)	23.7	21.8	16.7	17
Negative predictive value (%)	94.1	95.5	96.8	97.1

*This is the only scenario where the optimal threshold would be different when based on the maximum sum of sensitivity and specificity or on a minimal difference between sensitivity and specificity. In this case, the threshold was chosen based on the maximum sum of sensitivity and specificity.

antibiotic treatment at that point and may help avoid unnecessary courses of empiric broad-spectrum antibiotics that can sometimes be given for several days due to delays in the turnaround time of BC results.[28] When a specific infection such as pneumonia is very likely, then antibiotic treatment will be initiated regardless of the BC draw. However, in these cases our algorithm can still be used to withhold unnecessary BC testing.

Another clinically relevant aspect of this study is that we were able to show that routine laboratory results are associated with positive BCs. A low lymphocyte count appears to be related to a positive BC. This association has been described in earlier studies, but this variable has not been included in bacteraemia prediction models up until now.[29 30] Bilirubin is another notably strong predictor of a positive BC. Elevated bilirubin levels have been observed in patients with sepsis, and it is included in prognostic scores, such as the Sequential Organ Failure Assessment (SOFA) score for sepsis.[31 32] The association with positive BCs of other variables such as thrombocyte counts, temperature, blood pressure, heart rate and age is in line with previous studies.[2 24 33]

Strengths

The main strength that distinguishes this work from what has been done before is the comprehensive pipeline from raw data to model. The preprocessing and feature engineering phases were conducted in collaboration with a machine learning scale-up company (Pacmed, the Netherlands), which has considerable experience with machine learning in healthcare. The strategy towards the selection of features and algorithms that were used to predict BC outcomes presents a significant improvement over currently accepted methods in the medical literature since they provide a way to adapt the models to a specific hospital environment and thereby using the strengths of machine learning. Our pipeline used all available data so that the models themselves would decide on the importance of any feature.

With this approach, the models were not limited by the selection of features through current medical knowledge and had the potential to discover unknown associations with bacteraemia. Throughout preprocessing stages, we put emphasis on only using data that would routinely be available at the end of the ED visit, when the final treatment and admission decisions have to be made. This approach facilitates straightforward implementation of the models in clinical practice, without the need for additional data capture. Finally, we compare the results of the more complex gradient boosted tree model with a simpler logistic regression that is easier to understand for physicians, to improve the overall interpretability.

Limitations

There are several important limitations within this study. First, defining a positive BC is difficult. Our definition of contamination, which was defined as BCs that grew pathogens that are generally considered contaminants, is in line with previous literature.[2 4 12 13] However, we were not able to incorporate clinical characteristics when determining the positivity of the outcome, as is often done in practice. Therefore, it is still possible that samples that were mapped as contamination actually represented a true pathogen according to the operational definition in practice. However, the true positive rate of collected BC's in our population was somewhat higher than those described in previous literature.[4 6 33 34] This may be due to conservative mapping of pathogens to likely contaminants. A related limitation is that the model should not be used when a physician wants to detect a clinically relevant blood stream infection with pathogens that we considered to be contaminants, as with suspected central line-associated bloodstream infections (CLABSI). Our algorithm should be used as additive to the clinical pretest probability of bacteraemia, based on syndromes with a high likelihood of bacteraemia reported in earlier studies.[35]

Another limitation of this study is that various potentially predictive variables could not be adequately extracted from the EHR system. Comorbidities, medication at home and placement of lines are not well documented within the EHR and this data would not be reliable enough to use in a prediction model. Furthermore, we were not able to use free-text data due to privacy concerns. Therefore, we could not use physician and nurse reports.

A final important limitation is that this study is performed in a single-centre setting and external validation of the models is necessary. Not all variables will be available in each hospital worldwide due to heterogeneity between healthcare systems. A strength of using machine learning algorithms in clinical practice, as opposed to static and general risk scores such as the Shapiro decision rule, is that they can adapt to the local situation and change over time. However, to maintain this advantage, a dedicated effort to use our extensive data pipeline in each individual hospital is necessary in order to adapt to the local situation. This requires a considerable time investment.

Future research

Our current study gives rise to several potential follow-up studies. First, external validation is a key aspect to ensure that we find a true signal of positive BCs and that there is little overfitting to confounding factors in our single centre. External validation of the exact algorithm we used in our hospital is hard, since all variables need to be measured in the other centre as well. Therefore, there are two main options for external validation. We either need to use our complete pipeline and create a modified model which we test specifically in a different centre. Or else, we will need to simplify the current model and only select the most important and generally measured features, so that the exact model can be tested in other settings. Furthermore, we also need to prospectively validate the findings through an integration of the model into the realtime EHR environment before we come to an intervention study with our model. This way, we can observe whether the model performance remains stable over time or whether systematic retraining protocols are needed.

Additionally, there is a need to further explore variables that are highly associated with positive BCs. If we start measuring such factors in clinical practice, then they can easily be incorporated in the algorithms. For example, various studies have shown that procalcitonin can predict BC positivity with good performance.[19 36] We would be very interested to see the performance of a model created based on our pipeline in a hospital that regularly measures procalcitonin, as this may improve the performance substantially. Another important step could be to include additional clinical information by using free-text data.

In conclusion, we created two models that predict BC outcomes in the ED with AUROCs of 0.77 (95% CI 0.73 to 0.82) and 0.78 (95% CI 0.73 to 0.82) in this single-centre setting. The models are based on routinely captured clinical data and are therefore well suited for implementation in clinical practice. Further research is necessary for external and prospective validation of the models and implementation studies to identify potential benefits and possible risks. The main value of these models lies in the ability to identify patients at low risk of bacteraemia, which can help reduce unnecessary BC testing and provides an additional tool to decide whether antibiotic treatment is needed. Based on the model predictions, we would be able to withhold BC testing in 70% of the population with few omission.

Author affiliations

[1]Section General and Acute Internal Medicine, Department of Internal Medicine, Amsterdam Public Health Research Institute, Amsterdam UMC Location VUmc, Amsterdam, The Netherlands
[2]Department of Clinical Chemistry, Amsterdam UMC Location VUmc, Amsterdam, The Netherlands
[3]Center for Experimental and Molecular Medicine, Amsterdam UMC Location AMC, Amsterdam, The Netherlands
[4]Pacmed, Amsterdam, The Netherlands
[5]Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam Cardiovascular Science, Amsterdam Infection and Immunity Institute, Amsterdam UMC Location VUmc, Amsterdam, The Netherlands
[6]Section Infectious Diseases, Department of Internal Medicine, Amsterdam UMC Location AMC, Amsterdam, The Netherlands
[7]Board of Directors, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands

Contributors AWB and MS contributed equally to this paper. PWBN was the principal investigator and guarantor of the study. AWB, MS, LCAP and PWBN designed the study. ESvdE, LCAP, AYvdZ, PWGE, MHHK, AWB, MS and PWBN contributed to the acquisition of data. LM, JZ, MGS, AWB, MS, WJW and PWBN did the preparation of the data. LM, JZ and MGS were responsible for analysis of the data. AWB, MS, PWBN, TCM, PWGE and RdJ contributed to the interpretation of

the data. AWB, MS and LM drafted the first version of the manuscript. All authors revised the manuscript and approved the final version for publication.

Funding This project was funded by a research grant (no grant number available) from the Dutch federation for acute internal medicine (NVIAG).

Disclaimer The funder had no involvement in any part of the study.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study protocol was assessed by the local Medical Ethics Review Committee, which determined that a formal ethical review was not required according to the Dutch Medical Research Involving Human Subjects Act (WMO).(IRB number: IRB00002991; case: 2020.486). An appropriate registration with the local privacy officer was performed and the need for informed consent was waived. Only pseudonymised data were used in this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The data that support the findings of this study are available from the corresponding author on a reasonable request and when allowed by local privacy regulations.

ORCID iD

Prabath W B Nanayakkara http://orcid.org/0000-0002-1555-3682

REFERENCES

1 Wang HE, Jones AR, Donnelly JP. Revised national estimates of emergency department visits for sepsis in the United States*. *Crit Care Med* 2017;45:1443–9.

2 Coburn B, Morris AM, Tomlinson G, *et al*. Does this adult patient with suspected bacteremia require blood cultures? *JAMA* 2012;308:502–11.

3 Goto M, Al-Hasan MN. Overall burden of bloodstream infection and nosocomial bloodstream infection in North America and Europe. *Clin Microbiol Infect* 2013;19:501–9.

4 Nannan Panday RS, Wang S, van de Ven PM, *et al*. Evaluation of blood culture epidemiology and efficiency in a large European teaching hospital. *PLoS One* 2019;14:e0214052.

5 Jessen MK, Mackenhauer J, Hvass AMW. Prediction of bacteremia in the emergency department: an external validation of a clinical decision rule. *Eur J Emerg Med* 2016;23:44–9.

6 Denny KJ, Sweeny A, Crilly J, *et al*. Is it time for a culture change? blood culture collection in the emergency department. *Emerg Med Australas* 2018;30:575–7.

7 Bates DW, Goldman L, Lee TH. And resource utilization: the true consequences of false-positive results. *JAMA J Am Med Assoc* 1991;265:365–9.

8 Zwang O, Albert RK. Analysis of strategies to improve cost effectiveness of blood cultures. *J Hosp Med* 2006;1:272–6.

9 Dempsey C, Skoglund E, Muldrew KL, *et al*. Economic health care costs of blood culture contamination: a systematic review. *Am J Infect Control* 2019;47:963–7.

10 Schinkel M, Paranjape K, Nannan Panday RS, *et al*. Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput Biol Med* 2019;115:103488.

11 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.

12 Harvey DJ, Albert S. Standardized definition of contamination and evidence-based target necessary for high-quality blood culture contamination rate audit. *J Hosp Infect* 2013;83:265–6.

13 Dargère S, Cormier H, Verdon R. Contaminants in blood cultures: importance, implications, interpretation and prevention. *Clin Microbiol Infect* 2018;24:964–9.

14 Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol* 2019;34:23–36.

15 Jiménez-Valverde A, Lobo JM. Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecol* 2007;31:361–9.

16 Pewsner D, Battaglia M, Minder C, *et al*. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. *BMJ* 2004;329:209–13.

17 5.10 SHAP (SHapley additive exPlanations), interpretable machine learning. interpret. Mach. learn. Available: https://christophm.github.io/interpretable-ml-book/shap.html

18 Eliakim-Raz N, Bates DW, Leibovici L. Predicting bacteraemia in validated models--a systematic review. *Clin Microbiol Infect* 2015;21:295–301.

19 Ratzinger F, Haslacher H, Perkmann T, *et al*. Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study. *Sci Rep* 2018;8:12233.

20 Van Steenkiste T, Ruyssinck J, De Baets L, *et al*. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. *Artif Intell Med* 2019;97:38–43.

21 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* 2020;46:454–62.

22 Liu VX, Wiens J. 'No growth to date'? predicting positive blood cultures in critical illness. *Intensive Care Med* 2020;46:525–7.

23 Charani E, Ahmad R, Rawson TM, *et al*. The differences in antibiotic decision-making between acute surgical and acute medical teams: an ethnographic study of culture and team dynamics. *Clin Infect Dis* 2019;69:12–20.

24 Shapiro NI, Wolfe RE, Wright SB, *et al*. Who needs a blood culture? A prospectively derived and validated prediction rule. *J Emerg Med* 2008;35:255–64.

25 Tudela P, Lacoma A, Prat C, *et al*. Predicción de bacteriemia en Los pacientes Con sospecha de infección en urgencias. *Medicina Clínica* 2010;135:685–90.

26 Pawlowicz A, Holland C, Zou B, *et al*. Implementation of an evidence- based algorithm reduces blood culture overuse in an adult emergency department. *Gen Int Med Clin Innov* 2016;1:26–9.

27 Denny KJ, Gartside JG, Alcorn K, *et al*. Appropriateness of antibiotic prescribing in the emergency department. *J Antimicrob Chemother* 2019;74:515–20.

28 Sweeney TE, Liesenfeld O, May L. Diagnosis of bacterial sepsis: why are tests for bacteremia not sufficient? *Expert Rev Mol Diagn* 2019;19:959–62.

29 de Jager CPC, van Wijk PTL, Mathoera RB, *et al*. Lymphocytopenia and neutrophil-lymphocyte count ratio predict bacteremia better than conventional infection markers in an emergency care unit. *Crit Care* 2010;14:R192.

30 Laukemann S, Kasper N, Kulkarni P, *et al*. Can we reduce negative blood cultures with clinical scores and blood markers? results from an observational cohort study. *Medicine* 2015;94:e2264.

31 Singer M, Deutschman CS, Seymour CW, *et al*. The third International consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016;315:801.

32 Minemura M, Tajiri K, Shimizu Y. Liver involvement in systemic infection. *World J Hepatol* 2014;6:632–42.

33 Jessen MK, Mackenhauer J, Hvass AMSW, *et al*. Prediction of bacteremia in the emergency department. *Eur J Emerg Med* 2016;23:44–9.

34 Bates DW, Cook EF, Goldman L, *et al*. Predicting bacteremia in hospitalized patients. A prospectively validated model. *Ann Intern Med* 1990;113:495–500.

35 Fabre V, Sharara SL, Salinas AB, *et al*. Does this patient need blood cultures? A scoping review of indications for blood cultures in adult nonneutropenic inpatients. *Clin Infect Dis* 2020;71:1339–47.

36 Chirouze C, Schuhmacher H, Rabaud C, *et al*. Low serum procalcitonin level accurately predicts the absence of bacteremia in adult patients with acute fever. *Clin Infect Dis* 2002;35:156–61.