Utilising routinely collected clinical data through time series deep learning to improve identification of bacterial bloodstream infections: a retrospective cohort study

Damien K Ming, Vasin Vasikasin, Timothy M Rawson, Pantelis Georgiou, Frances J Davies, Alison H Holmes, Bernard Hernandez**

Summary

Background Blood cultures are the gold standard for diagnosing bacterial bloodstream infections, but test results are only available 24–48 h after sampling. We aimed to develop and evaluate models using health-care data to predict bloodstream infections in patients admitted to hospital.

Methods In this retrospective cohort study, we used routinely collected blood biomarkers and demographic data from patients who underwent blood sample collection for testing via culture between March 3, 2014, and Dec 1, 2021, at Imperial College Healthcare NHS Trust (London, UK) as model features. Data up to 14 days before blood sample collection were provided to long short-term memory (LSTM) or static logistic regression models. The primary outcome was prediction of blood culture results, defined as a pathogenic bloodstream infection (ie, isolation of pathogenic bacteria of interest) or no bloodstream infection (ie, no growth or contamination). Data collected up to Feb 28, 2021 (n=15 212) comprised the training set and were evaluated against a temporal hold-out test set comprising patients who were sampled after March 1, 2021 (n=5638).

Findings Among 20 850 patients with available data, pathogenic bacteria were observed in the cultured blood samples of 3866 (18·5%) patients. 2920 (62·2%) of 4897 patients who had their blood samples taken more than 48 h after admission to hospital had pathogenic bloodstream infections, and so were defined as having hospital-acquired bloodstream infections. Including data from the 7 days before admission (7-day window approach) and using five-fold cross validation in the training set gave an area under receiver operator curve (AUROC) of 0·75 (IQR 0·68–0·82) and an area under the precision recall curve (AUPRC) of 0·58 (0·46–0·77) for static models and an AUROC of 0·92 (0·91–0·93) and AUPRC of 0·75 (0·72–0·76) for the LSTM model. In the hold-out test set performances were: AUROC of 0·74 (95% CI 0·70–0·78) and AUPRC of 0·48 (0·43–0·53) for static models and AUROC of 0·97 (0·96–0·97) and AUPRC of 0·65 (0·60–0·70) for LSTM. Removal of time series information resulted in lower model performance, particularly for hospital-acquired bloodstream infections. Dynamics of C-reactive protein concentration, eosinophil count, and platelet count were important features for prediction of blood culture results.

Interpretation Deep learning models accounting for longitudinal changes could support individualised clinical decision making for patients at risk of bloodstream infections. Appropriate implementation into existing diagnostic pathways could enhance diagnostic stewardship and reduce unnecessary antimicrobial prescribing.

Funding UK Department of Health and Social Care, the National Institute for Health and Care Research, and the Wellcome Trust.

Lancet Digit Health 2025; 7: e205–15

*Contributed equally as senior authors

Centre for Antimicrobial Optimisation (D K Ming PhD, V Vasikasin PhD, T M Rawson PhD, Prof P Georgiou PhD, Prof A H Holmes FMedSci, B Hernandez PhD), **Healthcare Protection Research Unit in Healthcare Associated Infections** (T M Rawson, F J Davies PhD, Prof A H Holmes), **and Centre for Bio-inspired Technology** (Prof P Georgiou), **Imperial College London, London, UK; Department of Internal Medicine, Phramongkutklao Hospital and College of Medicine, Bangkok, Thailand** (V Vasikasin); **Imperial College Healthcare NHS Trust, London, UK** (F J Davies); **Department of Global Health and Infectious Diseases, University of Liverpool, Liverpool, UK** (Prof A H Holmes)

Correspondence to: Dr Bernard Hernandez, Centre for Antimicrobial Optimisation, Imperial College London, London W12 0NN, UK **b.hernandez-perez@imperial.ac.uk**

Introduction

Bacterial bloodstream infections are a substantial health-care burden. Bloodstream infections were attributable to 2·91 million (95% CI 1·74–4·53) deaths worldwide in 2019.[1] Prompt recognition and appropriate treatment of bloodstream infections is associated with improved patient outcomes and are a core priority of many health-care guidelines.[2] Susceptible patient groups at high risk of bloodstream infection might benefit from early empirical antimicrobial therapy and enhanced monitoring. Conversely, identification of those at lower risk could support de-escalation of management and investigations or allow for antimicrobials to be safely withheld from patients pending investigation results. Such strategies must concurrently minimise selective pressure exerted to the individual and their surroundings to attenuate development of antimicrobial resistance.[3] The increasing incidence of nosocomial bloodstream infections,[4] coupled with increases in multi-drug resistant infections,[5] are contemporary challenges faced across health-care domains.

Definitive diagnosis of bloodstream infections is based on bacterial blood cultures. A low positive predictive value (PPV) of 7·5%,[6] result turnaround times of between 24 h and 48 h, and false positive results from skin flora contamination[7] lead to unnecessary testing and reduces

Research in context

Evidence before this study

We searched for publications on machine and deep learning approaches to predict bacterial bloodstream infections in a hospital setting on MEDLINE using medical subject headings and Google Scholar using terms: ("bacteraemia" OR "bloodstream infection" OR "BSI") AND "prediction" AND "artificial intelligence". Entries between Jan 1, 1994, and Jan 1, 2022, in English were included. Most studies used aggregated features over a specific time period to predict outcomes. Limitations across studies included unclear reporting of underlying methods (eg, the time series data handling), lack of independent cohort validation, and unclear microbiological definitions of outcome labels.

Added value of this study

We used a large cohort of patients undergoing blood sampling for cultures between 2014 and 2021, with no limitations on age or care setting across a UK National Health Service Trust in west London to develop and evaluate deep learning models. We found that through a time series approach, the use of 24 routinely collected patient features allowed for prediction of subsequent bloodstream infections at the point of clinical evaluation to a high level of discrimination. We demonstrate the specific contribution of longitudinal information in deep learning models with specific benefits for patients with hospital-acquired bloodstream infections.

Implications of all the available evidence

The consideration of patient state over time is crucial in clinical practice for patient evaluation. We found that explicit inclusion of these longitudinal data in deep learning models is of benefit in predicting bloodstream infection. If incorporated at point of care alongside laboratory diagnostics, these models could improve patient care through supporting antimicrobial and diagnostic stewardship interventions.

See Online for appendix

clinical usefulness (appendix pp 2–3). Considerable variations in practice during the pre-analytical phase contribute to under-ascertainment.[8] These factors are important because datasets generated from accurate bloodstream infections diagnoses are central to antimicrobial resistance surveillance and influence policy globally.[9] These priorities haven been emphasised by national bodies[10] but wider issues relating to the judicious use of diagnostics—namely, diagnostic stewardship—are equally crucial.[11] With appropriate implementation, such measures can reduce antimicrobial consumption,[12] improve health-care cost effectiveness,[13] and optimise allocation of resources, particularly in settings with limited capacity.[14]

Clinical scoring systems that use routinely collected clinical data could augment the diagnosis of bloodstream infections to improve the recognition in a cost-effective way.[15] Syndromes associated with bloodstream infections are inherently dynamic in nature and change over hours or days. However, most bedside scores and machine learning algorithms designed for the prediction of the likelihood of bloodstream infection while in hospital exclusively use feature information from a single timepoint and provide static approaches to patient evaluation.[16–18] The nature of these changes and their association with risk of bloodstream infection are not well understood. Existing research into the prediction of bloodstream infections is limited by small sample sizes, unclear microbiological definitions, and data methods (eg, lack of cross validation, hold-out test sets, or external validation; appendix pp 4–6). Therefore, we aimed to investigate the use of longitudinal patient information for the prediction of bloodstream infections and to better understand the added contribution of such time series data. Additionally, we analysed hospital-onset bloodstream infections, which are associated with increased delays in diagnosis and treatment.[19]

Methods

Study design

In this retrospective cohort study, we aimed to develop and evaluate a set of predictive models using data from the Imperial College Healthcare NHS Trust (London, UK) that can be applied at the point of blood sample acquisition for culturing to stratify the risk of pathogenic bloodstream infection in patients admitted to and staying in hospital. The wider goal of the study is the incorporation of these models into a clinical tool to support decision making in real time and augment identification of at-risk groups and subsequent management at point of care.

Prediction outcomes were binary and classified into either pathogenic bloodstream infection or not bloodstream infection states on the basis of subsequent blood culture outcomes for the individual. Using the WHO priority list of pathogens,[20] a pathogenic bloodstream infection was defined as the isolation of one or more of the following: *Escherichia coli*, *Klebsiella* spp, *Enterococcus* spp, *Pseudomonas* spp, *Proteus* spp, *Serratia* spp, *Citrobacter* spp, *Streptococcus* spp, and *Staphylococcus aureus*. Not bloodstream infections were defined as blood cultures that resulted in either no bacterial growth after 5 days of standard incubation, or isolation of only one or more of: coagulase-negative *Staphylococcus* group, *Micrococcus* spp, or *Corynebacterium* spp excluding *Corynebacterium striatum*. Inclusion of these bacterial species captures 84·5% of the overall positive blood cultures seen at our institution (Imperial College Healthcare NHS Trust, London, UK). Patients with blood cultures from which yeasts and other

minor bacterial species were isolated were excluded from analyses.

This is a retrospective analysis of de-identified, pseudonymised data from patients admitted to Imperial College Healthcare NHS Trust, London, UK, and individual patient consent was not required. Ethics approval for the use of these data was granted by London-Chelsea Research and Ethics Committee (reference 17/LO/0047; IRAS 204949) and by the Imperial Clinical Analytics, Research and Evaluation (iCARE) committee, in turn provided with approval by the South West–Central Bristol Research and Ethics Committee (reference 21/SW/0120; IRAS 282093). The TRIPOD checklist, and further details of the methods, are presented in the appendix (pp 7–14).

Patients and data sources

Datasets for development and evaluation of all models were derived from electronic health records at Imperial College Healthcare NHS Trust. The hospital trust consists of a group of teaching hospitals across three main locations in west London (Hammersmith, Charing Cross, and St Mary's Hospital in Paddington) with a centralised microbiology laboratory covering a population of 1·5 million people. Clinical and laboratory data were obtained from the Imperial Clinical Analytics, Research and Evaluation (iCARE) platform and the electronic healthcare records of the Trust on Sept 8, 2023.

Patients of all ages and care settings who underwent blood culture sampling at an Imperial College Healthcare NHS Trust site between March 3, 2014, and Dec 1, 2021, were included in the study. Two separate datasets were concatenated to maximise generalisability of findings. Dataset 1 comprised patients with a positive blood culture, irrespective of final organism species, who had their blood sample taken, tested, and results made available between March 3, 2014, and Dec 31, 2019, and which include both pathogenic and non-pathogenic species. Dataset 2 comprised all patients who underwent blood culture sampling at Imperial College Healthcare NHS Trust between Jan 1, 2020, and Dec 1, 2021, and included pathogenic species, non-pathogenic species, and samples that resulted in no growth.

Blood cultures are commonly collected in paired bottles which then undergo both aerobic and anaerobic incubation. Patients for whom the same bacterial species were isolated from both aerobic and anaerobic bottles were combined to comprise a single observation. For those with multiple positive cultures during the same hospital episode, only the first pathogenic result was considered. Patients who had blood cultures taken more than 48 h after hospital admission and who were found to have a pathogenic bloodstream infection were defined as having a hospital-acquired infection; if patients had samples taken 48 h or earlier after hospital admission and were found to have pathogenic infection, they were defined as having a community-acquired infection.

Data analysis

Patient age, sex, and blood biomarkers collected during standard clinical care up to 14 days preceding the date of blood sample collection where available were used as features in the model. Features were chosen on the basis of data availability, biological plausibility, and results from an iterative selection process in which different panel combinations underwent five-fold cross validation and evaluation for predictive performance. The final set of 24 features that was used across all models was derived from three blood biomarker panels (full blood count, biochemistry, and coagulation) and included patient demographics and point-of-care results from bedside analysers.

The whole dataset was split temporally into a training and validation set (75%) and a hold-out test set (25%), with the hold-out test set used exclusively only for model evaluation (figure 1A). Patients who underwent blood culture sampling between March 3, 2014 and Feb 28, 2021, inclusive (n=15 212) were included in the training set (and validation set where appropriate [ie, during cross validation]); patients sampled between March 1, 2021, and Dec 1, 2021 (n=5638) were separated out as the hold-out test set. Feature data were formatted using 24 h as the unit of time increment and mean values were calculated for measurements that were repeated more than once in the same day. The timepoint of blood sample collections was defined as day 0 (figure 1B). Feature data were scaled to a set range (between –1 and 1) after fitting to the training set only.

For more on the **iCARE platform** see https://imperialbrc.nihr.ac.uk/facilities/icare/

Baseline models were constructed via logistic regression using data from a feature window preceding blood culture acquisition (figure 1B). The last available value for each biomarker was used to emulate a static decision-making process at the point of clinical evaluation (ie, at blood culture acquisition [day 0]). A range of feature engineering approaches was used, including the use of summary feature values over the window period (mean, minimum, and maximum), inclusion of biomarker timepoints as covariates, and dimensionality reduction done through principal component analysis.

For time series models, recurrent neural network-based approaches were chosen as the primary approach to handle longitudinal data. Long short-term memory (LSTM) deep learning networks can explicitly model sequential time series data and consider long-term dependencies.[21] Therefore, the LSTM method was chosen as an exemplar of a standard deep learning approach that can handle sequential information without substantial feature engineering requirements. Models were trained and optimised through a Bayesian optimisation framework in which models constructed according to a hyperparameter grid were iteratively chosen depending on their performance.

Logistic regression and LSTM models underwent five-fold stratified random cross validation using the

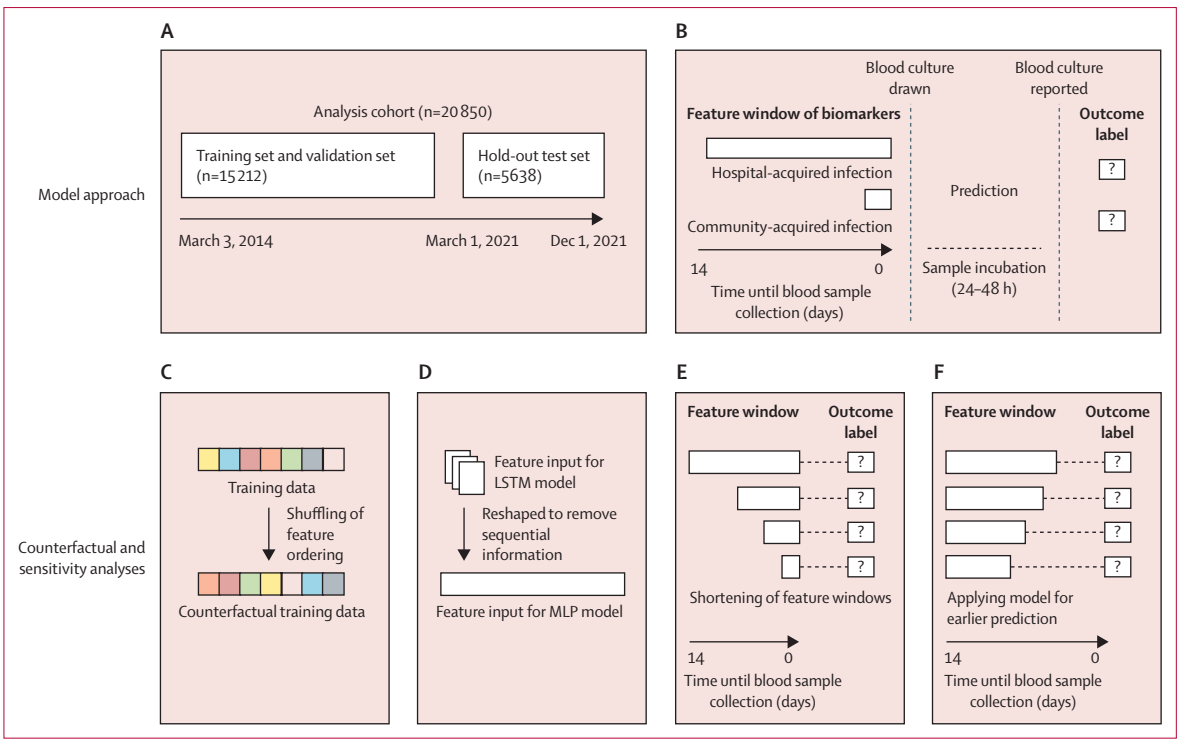


Figure 1: **Overview of model approach and counterfactual analyses**

(A) The temporal splitting of the analysis cohort into the training and hold-out test set. (B) Biomarker features collected up to 14 days before the day of blood culture draw were applied at timepoint 0 (day 0), and used for prediction of outcome labels. The clinical outcomes, provided by the microbiology laboratory, were available with a typical delay of 24–48 h after blood sample collection. (C) Sequential data removal through shuffling of biomarker order and applied in a LSTM model. (D) Sequential data removal through reshaping and used in a MLP model. Feature windows progressively shortened (E) and applied earlier (F) to evaluate the effect on predictive performance. LSTM=long short-term memory. MLP=multilayer perceptron.

training set alone and were evaluated against the hold-out test set. Outcome metrics were expressed in the form of area under the receiver operator curve (AUROC), area under the precision recall curve (AUPRC), sensitivity, specificity, PPV, negative predictive value (NPV), and Brier score (which measures the accuracy of probabilistic predictions, with lower values indicating better performance). Model metrics were calculated as median (IQR) when obtained in five-fold cross validation. Class imbalance in training was 22%, which was considered to not be meaningful, but experiments were performed to account for imbalance through changing the initial bias value. However, these experiments did not result in notable differences in performance. We used bootstrapping predictor–label pairs repeated 2000 times to derive 95% CIs to compare performance between models. We also evaluated other deep learning models that consider time series information, which were transformer, gated recurrent unit, and convolutional neural network.

Around 25% of all features were available at day 0, with the proportion of available data decreasing with earlier timesteps to around 14% at 7 days before blood sample collection. Feature data were more complete for full blood count and biochemistry data compared with other features. From initial data exploration, we found that the pattern of missing values was not at random but related to blood culture outcomes and differed between community-acquired and hospital-acquired infections. Missing data related to biomarker feature were more often observed for observations classified as hospital-acquired infections. Missing data were initially forward filled in time using a last observation carried forward approach when possible for both the training and hold-out test sets. For LSTM models, all biomarker sequences were padded to equal lengths and remaining missing values were masked using zero imputation and not considered in model training in deep learning models. Mean or median imputation was trialled for logistic regression models.

We did several post-hoc sensitivity and counterfactual analyses. To understand the contribution of the time series, we shuffled the overall ordering of biomarker features in every window in the training set to create a series of counterfactual datasets (figure 1C). This process altered temporal relationships between features while retaining absolute values, with the null hypothesis being that sequential relationships were unimportant and hence not exploited by sequential models to predict the likelihood of bloodstream infection. Models trained on

the counterfactual dataset were evaluated against an unchanged hold-out test set. The process was repeated ten times, each time using a different shuffled training set. We chose to repeat the process ten times as a balance between computational costs and ensuring stability, robustness, and generalisation of our findings. Separately, the effect of different prediction window sizes (14, 7, 3, and 1 day) on model performance were assessed for the LSTM model (figure 1E). Similarly, the predictions were applied earlier to evaluate the effect on predictive performance, also resulting in shortened windows (figure 1F). Deep learning approaches do not account for the sequential nature of data (eg, multilayer perceptron [MLP]) were also used as comparators (figure 1D). We examined whether prediction of blood culture outcomes to a Gram or species level was possible through formulating the problem as a multi-class problem using a LSTM approach. We also took a one-versus-rest approach for multi-class classification, which involves training a separate binary classifier for each class. Each classifier was trained to distinguish one specific class from all the others, treating the target class as positive and all other classes as negative. During prediction, the classifier that outputs the highest score is selected as the predicted class.

To assess interpretability and error, Shapley's additive values (SHAP values) for each feature by timestep were calculated using the DeepExplainer kernel[22] for all observations in the hold-out test set (n=5638) and a random sample (n=3000) from the training set. The SHAP values were used to estimate model feature importance. Error analyses were done through subset stratification and comparison of performances.

Statistical significance was determined from 95% CIs; however, for error analyses we defined significance as a p value of less then 0·0001.

Analyses were done using Python (version 3.7.1) and SciPy (1.15).

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or the writing of the report.

Results

We included 20 850 unique patients who underwent blood culture sampling (ie, sample acquisition, testing, and receipt of results) between March 3, 2014, and Dec 1, 2021. The median age of the patients was 58 years (IQR 33–74) and 10 178 (48·8%) of 20 850 were female and 10 672 (51·2%) were male; data on race and ethnicity were not assessed. In total, bacterial isolates for the full dataset included coagulase negative *Staphylococcus* (2465 [11·8%] of 20 850, *E coli* (1491 [7·2%], *Streptococcus* spp (615 [2·9%]), *S aureus* (473 [2·3%]), *Klebsiella* spp (386 [1·9%]), *Enterococcus* spp (326 [1·6%]) and others (575 [2·8%]; table 1). 3866 (18·5%) of 20 850 of isolates were classified as pathogenic bloodstream infections. 4697 (22·5%) of 20 850 patients underwent blood sample collection more than 48 h after hospital admission. Patients in this subset who had pathogenic bacteria isolated were defined as hospital-acquired bloodstream infection (2920 [62·2%] of 4897). Baseline characteristics and microbiological outcomes of patient groups are shown in table 1. There were differences in the distribution of missing values between pathogenic and non-pathogenic culture observations (appendix pp 11–12).

	Full cohort (N=20 850)	Training set (n=15 212)	Hold-out test set (n=5638)
Age, years	58 (33–74)	60 (36–76)	49 (26–69)
Sex			
Female	10 178 (48·8%)	7353 (48·3%)	2813 (49·9%)
Male	10 672 (51·2%)	7859 (51·7%)	2825 (50·1%)
Time of blood sample collection			
≤48 h after admission	16 153 (77·5%)	10 964 (72·1%)	5189 (92·0%)
>48 h after admission	4697 (22·5%)	4248 (27·9%)	449 (8·0%)
Microbiological outcomes			
Not bloodstream infection	16 984 (81·5%)	11 820 (77·7%)	5164 (91·6%)
No bacterial growth	14 316 (68·7%)	9536 (62·7%)	4780 (84·8%)
Coagulase negative *Staphylococcus*	2465 (11·8%)	2122 (13·9%)	343 (6·1%)
Others	203 (1·0%)	162 (1·1%)	41 (0·7%)
Pathogenic bloodstream infection	3866 (18·5%)	3392 (22·3%)	474 (8·4%)
Escherichia coli	1491 (7·2%)	1328 (8·7%)	163 (2·9%)
Streptococcus spp	615 (2·9%)	541 (3·6%)	74 (1·3%)
Staphylococcus aureus	473 (2·3%)	414 (2·7%)	59 (1·0%)
Klebsiella spp	386 (1·9%)	329 (2·2%)	57 (1·0%)
Enterococcus spp	326 (1·6%)	275 (1·8%)	51 (0·9%)
Others	575 (2·8%)	505 (3·3%)	70 (1·2%)
Hospital-acquired pathogenic bloodstream infection	2920/4697 (62·2%)	2677/4248 (63·0%)	243/449 (54·1%)
Community-acquired pathological bloodstream infection	946/16 153 (5·9%)	715/10 964 (6·5%)	231/5189 (4·5%)

Data are n (%), n/N (%), or median (IQR).

Table 1: **Baseline characteristics of the full analysis cohort, training set, and hold-out test set**

Logistic regression using the most recent biomarker features to the point of blood sample acquisition was used to establish baseline performance. Through five-fold cross validation on the training set (n=15 212), the model was associated with a median AUROC 0·75 (IQR 0·68–0·82) and AUPRC of 0·58 (0·46–0·77) for predicting the bloodstream infection state (table 2). Performance of this model on the hold-out test set (n=5638) yielded an AUROC of 0·74 (95% CI 0·70–0·78), AUPRC of 0·48 (0·43–0·53), PPV of 0·52, and NPV 0·96. Additional feature engineering approaches were evaluated, including (1) use of the mean and summary feature values over the window period, (2) inclusion of all features and timepoint of collection, and (3) dimensionality reduction through principal component

	Five-fold cross validation (training set; (IQR)		Hold-out test set performance (95% CI)						
	AUROC	AUPRC	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	Brier score
Logistic regression									
Most recent value	0·75 (0·68–0·82)	0·58 (0·46–0·77)	0·74 (0·70–0·78)	0·48 (0·43–0·53)	0·58	0·95	0·52	0·96	0·07
Mean value	0·73 (0·70–0·83)	0·55 (0·45–0·80)	0·71 (0·67–0·74)	0·45 (0·40–0·50)	0·52	0·96	0·53	0·96	0·07
Summarised	0·72 (0·67–0·87)	0·58 (0·46–0·77)	0·71 (0·67–0·74)	0·43 (0·38–0·49)	0·52	0·96	0·55	0·96	0·07
All features and timepoint	0·72 (0·59–0·79)	0·78 (0·45–0·85)	0·61 (0·56–0·65)	0·46 (0·41–0·52)	0·60	0·96	0·64	0·96	0·09
PCA (three component)	0·78 (0·76–0·90)	0·63 (0·50–0·90)	0·72 (0·69–0·75)	0·47 (0·41–0·51)	0·71	0·96	0·61	0·97	0·09
LSTM									
7-day window	0·92 (0·92–0·93)	0·75 (0·72–0·76)	0·97 (0·96–0·97)	0·65 (0·60–0·70)	0·93	0·98	0·56	1·00	0·04
14-day window	0·92 (0·91–0·92)	0·74 (0·71–0·74)	0·97 (0·96–0·97)	0·65 (0·60–0·70)	0·93	0·98	0·56	1·00	0·04

A five-fold random cross validation strategy on training data alone was used to generate median (IQR) estimates. Hold-out test set performances, including sensitivity, specificity, PPV, NPV, and Brier score are shown for each approach. The mean and summarised feature engineering approaches for the logistic regression were done without imputation. PCA was done after biomarker values were summarised. AUROC=area under the receiver operator curve. AUPRC=area under the precision recall curve. PCA=principal component analysis. LSTM=long short-term memory. PPV=positive predictive value. NPV=negative predictive value.

Table 2: **Summary of model results for logistic regression and LSTM model approaches for predicting bloodstream infection state**

analysis. These approaches were not associated with significant differences in model performance compared with baseline performance.

The LSTM model was trained and evaluated using the same data. Biomarker features from either a 7-day or 14-day window preceding the date of blood sample collection were included. Through five-fold cross validation of the training set, the median AUROC was 0·92 (IQR 0·92–0·93) with a 7-day window and 0·92 (0·91–0·92) with a 14-day window, and the median AUPRC was 0·75 (0·72–0·76) with a 7-day window and 0·74 (0·71–0·74) with a 14-day window (table 2). Performance on the entire hold-out test set was associated with an AUROC of 0·97 (95% CI 0·96–0·97) with a 7-day window and 0·97 (0·96–0·97) with a 14-day window, an AUPRC of 0·65 (0·60–0·70) with a 7-day window and 0·65 (0·60–0·70) with a 14-day window, with a PPV of 0·56 and NPV of 1·00 with the 7-day window and with a 14-day window (table 2). In a post-hoc sensitivity analysis, we performed a rolling window evaluation over the hold-out test set period (March 1 to Dec 1, 2021). The trained LSTM model was tested against observations from a 2-month window, and then rolled forward in 1-monthly steps. Discriminant performances for observations in each of the windows were similar to overall performance, with ranges of AUROC of 0·70–0·98 and AUPRC of 0·61–0·75 (appendix p 15).

Model performances in patients admitted to hospital more than 48 h before acquisition of blood sample were specifically examined because of the clinical importance of this cohort. Patients in this cohort are often more frail and likely to experience complications, which results in a worse clinical outcome. 2920 (62·2%) of 4697 patients in this group had a pathogenic bloodstream infection and were defined as having hospital-acquired bloodstream infections. A LSTM model trained on this subset had higher AUROC and AUPRC estimates than with a logistic regression approach (table 3). However, we found no significant differences between the models when assessed in the hold-out test set (table 3).

To examine the importance of time series information, models were trained on a set of counterfactual training data. This was done by shuffling the ordering of features for each 7-day window (eg, features collected from each timepoint were replaced with those from another; figure 1C). Therefore, feature values remained intact but were presented to the model in a different order. Median performance of these models when tested against the unshuffled hold-out test set, repeated ten times, was associated with an AUROC of 0·62 (IQR 0·46–0·84) and AUPRC of 0·28 (0·16–0·35). The use of a multilayer perceptron model (which does not handle sequential data) was associated with lower performance, but only within the hospital-acquired infection cohort (figure 1D; appendix p 20).

We shortened the feature window length to evaluate the minimum duration of time series data required for prediction (figure 1E). A 3-day feature window length was associated with similar performance to models trained against 7-day or 14-day models (table 2); however, shorter window lengths were associated with lower performance when applied in the hospital-acquired infection cohort (appendix p 19).

Considering delays in bloodstream infection diagnosis and blood sample acquisition in clinical care, we examined the ability of the LSTM model to make predictions when applied at timepoints earlier than date of blood sample acquisition (figure 1F). Feature windows were progressively shortened by removal of days 0 to −3. Discriminant performance of the model through cross validation decreased with shorter time windows, but AUPRC values when the model was applied 1 day before the point of blood sample acquisition were similar (appendix p 17).

The inclusion of bacterial species and Gram stain information in the LSTM model as well as pathogenicity through a multi-class approach resulted in similar

	Five-fold cross validation (training set; IQR)		Hold-out test set performance (95% CI)						
	AUROC	AUPRC	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	Brier score
Logistic regression									
Most recent value	0·67 (0·67–0·68)	0·77 (0·76–0·78)	0·62 (0·57–0·67)	0·67 (0·61–0·73)	0·48	0·71	0·66	0·52	0·24
LSTM									
7-day window	0·75 (0·74–0·76)	0·83 (0·82–0·84)	0·64 (0·58–0·69)	0·66 (0·58–0·72)	0·54	0·71	0·69	0·55	0·24
14-day window	0·74 (0·73–0·74)	0·81 (0·78–0·82)	0·65 (0·60–0·70)	0·69 (0·63–0·75)	0·59	0·66	0·68	0·57	0·24

Metrics are derived from five-fold cross validation through training data only, and evaluated against the hold out test set. AUROC=area under the receiver operator curve. AUPRC=area under the precision recall curve. PCA=principal component analysis. LSTM=long short-term memory. PPV=positive predictive value. NPV=negative predictive value.

Table 3: **Model performances using only data from patients with a hospital-acquired bloodstream infection**

median performance in the hold-out test set when testing against the baseline models (median AUROC using a one-versus-rest approach 0·97; AUPRC 0·73), where most of the classes represented non-pathogenic coagulase negative *Staphylococcus*. However, this approach also substantially exacerbated class imbalances between the training set and hold-out test set, leading to a significantly lower mean AUPRC of 0·20 and PPV (mean 0·14 across classes; data not shown). Other deep learning approaches that handle sequential data (transformer, convolutional neural network and gated recurrent unit models) were evaluated but these were not associated with differences in overall performance (data not shown).

In post-hoc analyses, feature importance derived from SHAP values were similar in the training set and hold-out test set. Increased C-reactive protein and a decreased eosinophil count and platelet count preceding blood sample acquisition were the most important features for the model and were associated with prediction of bloodstream infection state (figure 2). We found no substantial differences in model performance according to sex, but the model did perform significantly worse in patients younger than 1 year than in patients in all other age groups. The AUROC was also significantly lower for patients with a hospital-acquired bloodstream infection than in those with a community-acquired infection (table 4).

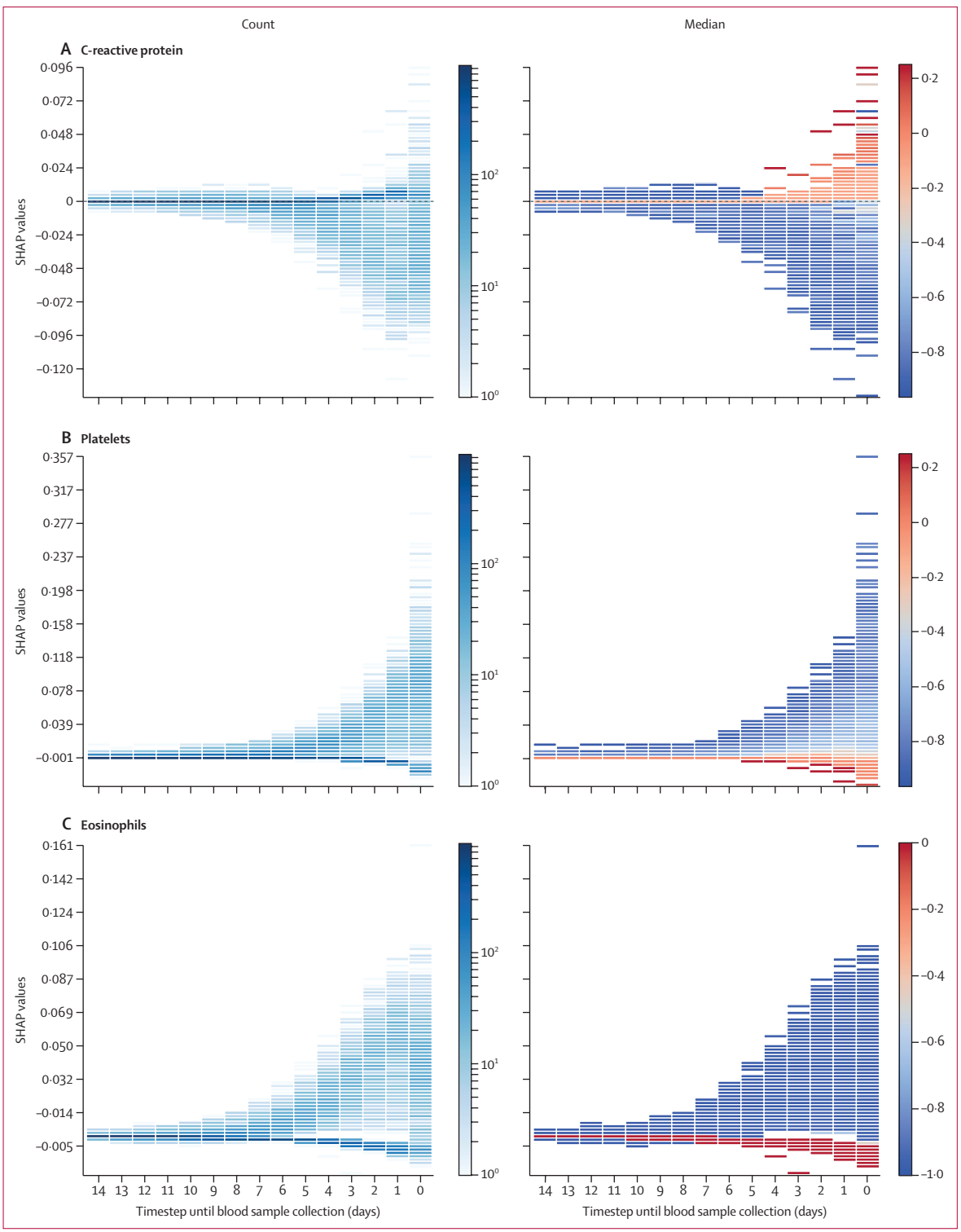
Discussion

Blood cultures are often under-utilised in clinical practice worldwide. Uptake is affected by a range of complex factors including staff training, clinical prioritisation, and access to diagnostics.[23] Interventions that enhance diagnostic stewardship through identification of individuals most in need of assessment can increase trust and value in blood cultures. The reduction in unnecessary testing is equally important, particularly for health-care settings with resource limitations and that have been exacerbated by global supply issues relating to blood culture media.[24] We developed and evaluated deep learning prediction models for early identification of bloodstream infection across a multi-site tertiary health-care centre within the UK NHS. Through use of routinely collected clinical information, models that explicitly accounted for the dynamic time series relationships between features offered greater predictive ability over time-invariant approaches. Removal of embedded sequential information through counterfactual and sensitivity analyses resulted in lower performance. The implementation of such models could facilitate individualised risk stratification at the point of clinical assessment, support treatment decisions in the 24-to-48 h period before blood culture results, and play a role in surveillance through improved ascertainment or modelling of missed bloodstream infections.

Raised C-reactive protein concentrations, eosinopenia, and thrombocytopenia were identified as important features for model performance, all of which have biological plausibility related to inflammatory response and their effects on the bone marrow—known parameters that change in severe bacterial infections and sepsis.[25] Changes in C-reactive protein SHAP values up to 4 days before blood sample acquisition were observed, and performance of the model when applied 1 day before blood sample acquisition was similar to that when applied at the date of blood sample acquisition (day 0), suggesting that earlier prediction of bloodstream infections might be possible. Prompt identification of bloodstream infection is crucial in improving patient outcomes,[26] particularly in hospitalised patients who are often more acutely unwell than those with community-acquired bloodstream infections and this hospitalised patient cohort typically has a higher rate of multi-drug resistant infections.[27] Although there is greater availability of longitudinal clinical data in patients who have been hospitalised for longer, lower performances might reflect the greater heterogeneity in these patients than in those with community-acquired bloodstream infections and a more difficult prediction task.[28]

Specific strengths of the study include a large, diverse patient cohort with no restriction on age and including patients admitted across many different care settings over the 7-year period, increasing generalisability of the results. Use of confirmed blood culture results as a clinically relevant and objective outcome is important for development of tools specifically for antimicrobial prescribing and infection management. The requirement for routinely collected data supports implementation and

also enables opportunities in automated bloodstream infection surveillance, which is important in understanding incidence of hospital-acquired infections.[29] The model was evaluated using a temporal split strategy, which increases robustness because the hold-out test period coincided with a period of high COVID-19 incidence in 2021. Clinical presentations of COVID-19 are often associated with an inflammatory syndrome that mimics bacterial co-infection.[30] During the COVID-19 pandemic, changes in blood culture practice, including



use of personal protective equipment (with a higher rate of contaminants) might explain the lower performance seen in the hold-out test set than was seen with cross validation in the training set. The high NPV (1·0) and calibration of the LSTM model supports its role as a rule-out test. Clinical tools that identify patients for whom empirical antimicrobial treatment can be safely withheld or de-escalated—particularly for hospitalised patients—would support antimicrobial stewardship through reducing antimicrobial resistance, adverse drug reactions, and incidence of *Clostridioides difficile* infections.

Our study has several limitations. Most patients (77·5%) in our dataset had blood cultures drawn on or within 48 h of hospital admission and few had complete biomarker data for the full 14 days or 7 days before blood samples were taken. We found that 3-day and 7-day feature windows were associated with similar performances, although this association was not seen for hospital-acquired infections, for which longitudinal data were more complete. We acknowledge that the time series approach is attenuated in current implementation, with LSTMs suited to handling significantly more complex sequential data. Experiments with different feature engineering approaches for linear models could yield similar results, although the aim of this study was not to compare these approaches, but rather to determine the effects of accounting for times series relationships. Feature data were missing for a substantial proportion of patients, with an average of 25% of features being recorded at the time of blood sample collection in our population. Our use of substantial masking and imputation could affect model stability. There were differences in the distribution of missing values between pathogenic and non-pathogenic culture observations and we speculate that the distribution of missing values was probably an important aspect in model performance that warrants further study.

We classified patients whose cultures resulted in the isolation of organisms of low pathogenicity (eg, the coagulase negative *Staphylococci* group) as having an outcome of no bloodstream infection in the absence of patient-level clinical data. Such bacteria can be pathogenic in specific settings, such as in neonatal populations, although the neonatal population represented a small proportion of the analysis population (data not shown). We also excluded yeast infections and other less common bacterial species with the assumption that these patients have a different set of risk factors. We concatenated two blood culture datasets from the same set of hospitals for analysis and had assumed that these were generally comparable given there was no substantial change in blood sample acquisition or culture methods over time. However, major differences between the datasets include the prevalence of COVID-19 co-infection, which would have only been present in dataset 2, and cultures that resulted in no growth were not reported in dataset 1, resulting in a difference in class imbalance between the training and hold-out test sets. Race and ethnicity data were not reported in the analysis because collection of

	AUROC (95% CI)	AUPRC (95% CI)
Sex		
Male	0·97 (0·97–0·98)	0·71 (0·65–0·78)
Female	0·98 (0·97–0·98)	0·69 (0·53–0·76)
Age group, years		
<1	0·98 (0·96–0·99)	0·34 (0·20–0·60)*
1 to <18	0·98 (0·96–0·99)	0·70 (0·51–0·85)
18 to <75	0·98 (0·97–0·98)	0·71 (0·66–0·77)
≥75	0·97 (0·96–0·98)	0·72 (0·64–0·81)
Bloodstream infection type		
Community acquired	0·99 (0·99–0·99)	0·76 (0·70–0·83)
Hospital acquired	0·64 (0·59–0·69)*	0·68 (0·62–0·74)
Care setting		
Medicine	0·97 (0·97–0·98)	0·67 (0·61–0·73)
Surgery	0·97 (0·94–0·98)	0·73 (0·61–0·84)
Haematology	0·95 (0·93–0·98)	0·76 (0·61–0·88)
Obstetrics	0·99 (0·99–1·00)	0·79 (0·63–0·95)
Paediatrics	0·97 (0·96–0·99)	0·57 (0·37–0·78)
Critical care	0·85 (0·72–0·96)	0·64 (0·35–0·88)
Nephrology	0·98 (0·95–0·99)	0·84 (0·71–0·97)

AUROC=area under the receiver operator curve. AUPRC=area under the precision recall curve. LSTM=long short-term memory. *Results that were significantly different (p<0·001) within the group.

Table 4: **Error analyses on the hold out test set for the LSTM 7-day model, stratified against available patient parameters**

Figure 2: **Post-hoc time series plots of a data availability histogram (left plot) and biomarker features with their measurement values (right plot) for C-reactive protein (A), platelets (B), and eosinophils (C)**

Since we are working with temporal data, each observation has dimensions F by T, where F represents the number of features (eg, biochemical markers) and T represents the number of time steps considered. In this figure, the x axis shows the days leading up to blood sample collection (day 0), and the y axis shows the SHAP values, which indicate the contribution of each feature to the model's predictions. Higher SHAP values correspond to greater contributions towards predicting the positive (pathogenic bloodstream infection) class. Because SHAP values are numeric, they have been binned for easier visualisation, with either the count (the numbers of times the value of the selected feature [eg, C-reactive protein] at time t (day –1) was available; the left plot) or median (the median value of the original feature value [eg, C-reactive protein value] for all the observations in which the SHAP value at time t fell within the bin; the right plot) used to colour the cells. The left plot, showing the count data, is a histogram representing data availability, indicating how often a value appeared in the observation (ie, not missing). Darker blue indicates greater availability. In the right plot, showing median data, the blue–red gradient shows the normalised feature value for each time step, with red indicating higher values and blue indicating lower ones. For example, panel A, as we approach the day of blood sample acquisition (day 0), higher values of C-reactive protein (indicated in dark red) are associated with higher SHAP values, contributing more to predicting the positive class. By contrast, lower values of C-reactive protein (indicated in dark blue) correspond to lower SHAP values, contributing to the negative class. On days further from blood sample acquisition (day 0), the SHAP values are lower, indicating that their contribution to the prediction is much smaller. SHAP values=Shapley's additive values.

this information was not harmonised across the two datasets used.

Finally, the absence of an alternative gold standard to blood cultures represents a ground truth problem. Blood cultures can be subject to variation between settings, including the clinical decision to obtain a sample, local laboratory processes, and data collection procedures. The diagnostic sensitivity of a single blood culture to diagnose bloodstream infection is also recognised to be suboptimal. Therefore, our study, which is based on data from only one group of health-care institutions, might limit the external generalisability of the results. True external validation of the model was not performed because publicly accessible datasets, such as Medical Information Mart for Intensive Care, capture a different health-care population and lack biomarker features that were important for our model, such as C-reactive protein. This is a limitation to our work; however, given heterogeneity in decision making and care pathways across diverse health-care settings, there are distinct advantages to the development and optimisation of local models. These factors relate to better model performance and limit issues relating to data sharing and privacy. Nonetheless, local models are inherently more prone to overfitting, which can affect interpretability. The finding that time series data are helpful for clinical prediction of bloodstream infections is likely to be broadly applicable to other health-care centres.

For more on **Medical Information Mart for Intensive Care** see https://mimic.mit.edu/

In conclusion, deep learning approaches using routinely collected biomarkers as features to stratify risk of bloodstream infection could complement use of existing blood culture diagnostics. The role of data as a diagnostic adjunct offers opportunities for automated surveillance of bloodstream infections. Clinical utility, patient outcomes, and cost-effectiveness will be the main measures of implementation success.[31] Carefully conducted and prospective evaluation that adheres to standardised methods will be important to realise these potential benefits.

Contributors

DKM and BH were the coordinators for the project and were responsible for the analyses, writing of the first draft of the manuscript, and had responsibility for the overall integrity of the dataset. DKM, BH, and VV accessed and verified the underlying data used in the study and performed the data curation and extraction through iCARE and additional analyses. PG provided supervision for the dataset curation and ethics application. TMR, FJD, and AHH provided clinical context to the analyses and discussion. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The code and example notebooks for the data analysis and the final trained LSTM model are available at a GitHub repository (hosted at https://github.com/bahp/tldigitalhealth-D-23-01381R1). Access to the iCARE dataset for individual level, biomarker, and microbiological data can be obtained through applying through https://www.imperial.ac.uk/medicine/research-and-impact/groups/icare/icare-facility/information-for-researchers/.

Acknowledgments

This research was funded in part by the Wellcome Trust CAMO-Net programme (grant reference 226691/Z/22/Z) at Imperial College London, the UK Department of Health and Social Care and the NIHR Imperial Biomedical Research Centre (NIHR203323). AHH is a senior NIHR investigator. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health and Social Care, NHS, or the NIHR. We are grateful to the Imperial College Healthcare NHS Trust and patients. This research was enabled by the iCARE secure data environment, and used the iCARE team and data resources.

References

1 GBD 2019 Antimicrobial resistance Collaborators. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2022; **400:** 2221–48.

2 Kadri SS, Lai YL, Warner S, et al. Inappropriate empirical antibiotic therapy for bloodstream infections based on discordant in-vitro susceptibilities: a retrospective cohort analysis of prevalence, predictors, and mortality risk in US hospitals. *Lancet Infect Dis* 2021; **21:** 241–51.

3 Holmes AH, Moore LSP, Sundsfjord A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* 2016; **387:** 176–87.

4 Goto M, Al-Hasan MN. Overall burden of bloodstream infection and nosocomial bloodstream infection in North America and Europe. *Clin Microbiol Infect* 2013; **19:** 501–09.

5 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022; **399:** 629–55.

6 Nannan Panday RS, Wang S, van de Ven PM, Hekker TAM, Alam N, Nanayakkara PWB. Evaluation of blood culture epidemiology and efficiency in a large European teaching hospital. *PLoS One* 2019; **14:** e0214052.

7 Hall KK, Lyman JA. Updated review of blood culture contamination. *Clin Microbiol Rev* 2006; **19:** 788–802.

8 Snyder JW. Blood cultures: the importance of meeting pre-analytical requirements in reducing contamination, optimizing sensitivity of detection, and clinical relevance. *Clin Microbiol Newsl* 2015; **37:** 53–57.

9 Lim C, Hantrakun V, Teerawattanasook N, et al. Impact of low blood culture usage on rates of antimicrobial resistance. *J Infect* 2021; **82:** 355–62.

10 NHS England. Improving the blood culture pathway. March 8, 2023. https://www.england.nhs.uk/wp-content/uploads/2022/06/B0686-improving-the-blood-culture-pathway-executive-summary-v1–1.pdf.pdf (accessed Sept 8, 2023).

11 Kleinert S, Horton R. Pathology and laboratory medicine: the Cinderella of health systems. *Lancet* 2018; **391:** 1872–73.

12 Woods-Hill CZ, Colantuoni EA, Koontz DW, et al. Association of diagnostic stewardship for blood cultures in critically ill children with culture rates, antibiotic use, and patient outcomes: results of the Bright STAR collaborative. *JAMA Pediatr* 2022; **176:** 690–98.

13 Dempsey C, Skoglund E, Muldrew KL, Garey KW. Economic health care costs of blood culture contamination: a systematic review. *Am J Infect Control* 2019; **47:** 963–67.

14 Iskandar K, Molinier L, Hallit S, et al. Surveillance of antimicrobial resistance in low- and middle-income countries: a scattered picture. *Antimicrob Resist Infect Control* 2021; **10:** 63.

15 Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect* 2020; **26:** 584–95.

16 Hernandez B, Ming DK, Rawson TM, et al. Advances in diagnosis and prognosis of bacteraemia, bloodstream infection, and sepsis using machine learning: a comprehensive living literature review. *Artif Intell Med* 2024; **160:** 103008.

17 Hernandez B, Herrero P, Rawson TM, et al. Supervised learning for infection risk inference using pathology data. *BMC Med Inform Decis Mak* 2017; **17:** 168.

18 Rawson TM, Hernandez B, Moore LSP, et al. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. *J Antimicrob Chemother* 2019; **74:** 1108–15.

19 Moehring RW, Sloane R, Chen LF, et al. Delays in appropriate antibiotic therapy for Gram-negative bloodstream infections: a multicenter, community hospital study. *PLoS One* 2013; **8:** e76225.

20 WHO. Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis. 2017. https://www.who.int/publications/i/item/WHO-EMP-IAU-2017.12 (accessed Oct 23, 3023).

21 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* 2020; **404:** 132306.

22 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30:** 4765–74.

23 Suntornsut P, Asadinia KS, Limato R, et al. Barriers and enablers to blood culture sampling in Indonesia, Thailand and Viet Nam: a Theoretical Domains Framework-based survey. *BMJ Open* 2022; **14:** e075526.

24 US Food and Drug Administration. Disruptions in availability of BD BACTEC blood culture media bottles—letter to health care providers. Aug 15, 2024. https://www.fda.gov/medical-devices/letters-health-care-providers/disruptions-availability-bd-bactec-blood-culture-media-bottles-letter-health-care-providers (accessed Oct 22, 2024).

25 Rawson TM, Mind D, Ahmad R, Moore LSP, Holmes AH. Antimicrobial use, drug-resistant infections and COVID-19. *Nat Rev Microbiol* 2020; **18:** 409–10.

26 Lodise TP, McKinnon PS, Swiderski L, Rybak MJ. Outcomes analysis of delayed antibiotic treatment for hospital-acquired *Staphylococcus aureus* bacteremia. *Clin Infect Dis* 2003; **36:** 1418–23.

27 Tabah A, Buetti N, Staiquly Q, et al. Epidemiology and outcomes of hospital-acquired bloodstream infections in intensive care unit patients: the EUROBACT-2 international cohort study. *Intensive Care Med* 2023; **49:** 178–90.

28 Lenz R, Leal JR, Church DL, Gregson DB, Ross T, Laupland KB. The distinct category of healthcare associated bloodstream infections. *BMC Infect Dis* 2012; **12:** 85.

29 Sips ME, Bonten MJM, van Mourik MSM. Automated surveillance of healthcare-associated infections: state of the art. *Curr Opin Infect Dis* 2017; **30:** 425–31.

30 Stringer D, Braude P. Myint PK, et al. The role of C-reactive protein as a prognostic marker in COVID-19. *Int J Epidemiol* 2021; **50:** 420–29.

31 Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. *NPJ Digit Med* 2022; **5:** 13.