TYPE Original Research
PUBLISHED 18 December 2023
DOI 10.3389/fmed.2023.1213411

Machine learning-based clinical decision support for infection risk prediction

Ting Feng[1], David P. Noren[1], Chaitanya Kulkarni[2], Sara Mariani[1], Claire Zhao[1], Erina Ghosh[1], Dennis Swearingen[3,4], Joseph Frassica[5], Daniel McFarlane[1] and Bryan Conroy[1]*

[1]Philips Research North America, Cambridge, MA, United States, [2]Philips Research Bangalore, Bengaluru, India, [3]Department of Medical Informatics, Banner Health, Phoenix, AZ, United States, [4]Department of Biomedical Informatics, University of Arizona College of Medicine, Phoenix, AZ, United States, [5]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States

OPEN ACCESS

EDITED BY
Francisco Martín-Rodríguez,
University of Valladolid, Spain

REVIEWED BY
Nuria Montes,
Hospital Universitario de La Princesa, Spain
Ulrich Bodenhofer,
University of Applied Sciences Upper Austria,
Austria

*CORRESPONDENCE
Bryan Conroy
✉ bryan.conroy@philips.com

RECEIVED 09 May 2023
ACCEPTED 21 November 2023
PUBLISHED 18 December 2023

CITATION
Feng T, Noren DP, Kulkarni C, Mariani S, Zhao C, Ghosh E, Swearingen D, Frassica J, McFarlane D and Conroy B (2023) Machine learning-based clinical decision support for infection risk prediction.
Front. Med. 10:1213411.
doi: 10.3389/fmed.2023.1213411

Background: Healthcare-associated infection (HAI) remains a significant risk for hospitalized patients and a challenging burden for the healthcare system. This study presents a clinical decision support tool that can be used in clinical workflows to proactively engage secondary assessments of pre-symptomatic and at-risk infection patients, thereby enabling earlier diagnosis and treatment.

Methods: This study applies machine learning, specifically ensemble-based boosted decision trees, on large retrospective hospital datasets to develop an infection risk score that predicts infection before obvious symptoms present. We extracted a stratified machine learning dataset of 36,782 healthcare-associated infection patients. The model leveraged vital signs, laboratory measurements and demographics to predict HAI before clinical suspicion, defined as the order of a microbiology test or administration of antibiotics.

Results: Our best performing infection risk model achieves a cross-validated AUC of 0.88 at 1 h before clinical suspicion and maintains an AUC >0.85 for 48 h before suspicion by aggregating information across demographics and a set of 163 vital signs and laboratory measurements. A second model trained on a reduced feature space comprising demographics and the 36 most frequently measured vital signs and laboratory measurements can still achieve an AUC of 0.86 at 1 h before clinical suspicion. These results compare favorably against using temperature alone and clinical rules such as the quick sequential organ failure assessment (qSOFA) score. Along with the performance results, we also provide an analysis of model interpretability via feature importance rankings.

Conclusion: The predictive model aggregates information from multiple physiological parameters such as vital signs and laboratory measurements to provide a continuous risk score of infection that can be deployed in hospitals to provide advance warning of patient deterioration.

KEYWORDS

healthcare-associated infection (HAI), machine learning, clinical decision support (CDS), model interpretability, pre-symptomatic infection risk

Background

Healthcare-associated infection (HAI), also referred to as nosocomial infection, remains a significant risk for hospitalized patients and a significant burden on healthcare systems. It has been reported that approximately 1 in 31 hospital patients develop an HAI on any given day (1), and nearly 99,000 people in the U.S. die annually from HAIs (2). Recent data shows that the incidence of HAI's increased during the pandemic (2020) revealing the fragile nature of interventions aimed at prevention (3). Over the last decade, the CDC has developed guidelines and strategies for the prevention of HAIs, focusing on improving clinical practice and antibiotic stewardship. While this guidance has shown some utility in lowering the incidence across several types of HAI, improving the outcomes for those who become infected remains challenging, particularly for the critically ill.

Early detection of *de-novo* infectious disease is critical for improving the outcomes of infected patients (4, 5), for the timely implementation of control measures critical to preventing its spread (6), and for reducing substantial healthcare costs associated with preventable HAIs (7). Hospitalized patients suffering from influenza, up to 20% of whom are nosocomial in origin, have better outcomes when treated with antiviral agents immediately after symptoms present (8). Antibiotic treatment has also been shown to be more effective in producing better outcomes for sepsis patients when administered early in the progression of the infection, particularly for mechanically ventilated patients (4, 5).

Clinical decision support (CDS) tools have received a great deal of attention over the last decade, including those focused on the detection of infection (9–11). Many of these CDS tools are rule based and developed through physician consensus and guidelines. These include more standardized solutions like the acute kidney injury (AKI) eAlert that has been deployed in hospitals in Wales (12, 13) and the National Early Warning Score (NEWS) that is standard for detecting general clinical deterioration in the United Kingdom (14). While these approaches benefit from clinician experience, they are simplified to remain generalizable and fail to capture the complete clinical context required to discriminate difficult or atypical cases. In addition, these approaches are not easily tailored or adapted, for example, to specific patient populations. More recently, several studies have suggested data-driven approaches to create physiological risk prediction algorithms, including in the areas of infection and sepsis prediction (9, 15–17).

This study uses machine learning applied on large retrospective hospital datasets to develop a clinical decision support (CDS) algorithm for the early detection of infection in hospitalized patients. By aggregating information across demographics and a set of 163 vital signs and laboratory measurements, we find our best-performing model can achieve a cross-validated AUC of 0.88 at 1 h before clinical suspicion and maintains an AUC >0.85 for the 48 h period prior to clinical suspicion of infection. By distilling the model down to a set of 36 most frequently measured vital signs, laboratory measurements and demographics, we can still maintain an AUC of 0.86 at 1 h before clinical suspicion. In the results, we further contrast our models against established clinical scoring systems—quick sequential organ failure assessment (qSOFA), and against tracking individual vital signs alone (e.g., temperature, etc.).

Methods

Description of data

We combined clinical data from three large hospital datasets: the MIMIC-III (Medical Information Mart for Intensive Care III) database comprising deidentified health-related data from patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (18), the eICU dataset from Philips' electronic ICU telemedicine business populated with deidentified patients' data from a combination of many critical care units throughout the continental United States between 2003 and 2016 (19), and a dataset of deidentified electronic medical records from patients who stayed in critical care units or low-acuity settings such as general wards in Banner Health collected from 2010 to 2015. In total, the combined dataset includes over 6.5 million patient encounters collected from more than 450 hospitals. Supplementary Figure S1 indicates the types of data present in each hospital dataset.

Ethical approval

The MIMIC-III project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Use of the eICU data was approved by the Philips Internal Committee for Biomedical Experiments. Banner Health data use was a part of an ongoing retrospective deterioration detection study approved by the Institutional Review Board of Banner Health and by the Philips Internal Committee for Biomedical Experiments. Requirement for individual patient consent was waived because the project did not impact clinical care, was no greater than minimal risk, and all protected health information was removed from the limited dataset used in this study.

Infection and control cohort extraction

We define infection patients as those who (1) have a confirmed infection diagnosis, and (2) have data indicating clinical suspicion of infection. Patients in the infection cohort were selected as those with confirmed infection diagnoses via ICD-9 and whose timing of clinical suspicion of infection could be localized by a microbiology culture test order. In the cases where more than one microbiology culture tests were ordered during the hospital stay, we used the earliest timing of the orders to mark clinical suspicion of infection for the given patient. Infection patients were then further screened into an HAI cohort if the timing of clinical suspicion of infection occurred at least 48 h after admission.

Patients in the control cohort were selected as those who have neither an infection-related ICD-9 diagnosis code nor any microbiology culture tests ordered. Since the selection criteria

Abbreviations: HAI, Healthcare-associated infection; CDS, Clinical decision support; qSOFA, Quick sequential organ failure assessment; SHAP, Shapley additive explanations; Spec, Specificity; TNR, True negative rate; AUC, Area under the ROC curve.

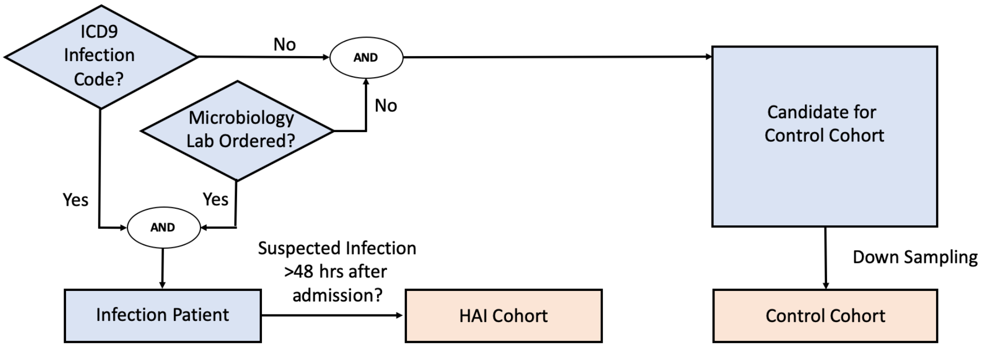
ICD9 Infection Code?
No
AND
Microbiology Lab Ordered?
No
Yes
Yes
AND
Candidate for Control Cohort
Down Sampling
Suspected Infection >48 hrs after admission?
Infection Patient
HAI Cohort
Control Cohort

FIGURE 1
Cohort inclusion/exclusion criteria flow diagram.

identified a much larger set of control patients than HAI patients, we randomly down-sampled the control cohort population without replacement to maintain a prior infection odds (prevalence) of 12.5%. This ensured that the training dataset would not be overly dominated by control patients, while still maintaining the HAI cohort as the minority class. Because our machine-learning methodology requires extracting clinical data before clinical suspicion of infection, we generated synthetic event times for the control patients, such that clinical data used for prediction for the control patients could be extracted in the same way as was done for the infection patients. To reduce bias, and to ensure sufficient data prior to event time for model building, we randomly assigned a time-point that is at least 48 h after the control patient's first clinical measurement, and that precedes the end of the control patient's hospital stay as the synthetic event time.

Figure 1 shows the general decision scheme for infection and control cohort extraction. Curation of infection ICD-9 codes is described in detail in the Supplementary material.

For a subset of eICU hospitals, due to limited availability of microbiology interfaces, microbiology charting data was either missing, sporadic, or incomplete. In such cases, the microbiology culture test criterion was replaced with the administration of non-prophylactic antibiotics. The cohort selection was otherwise the same: infection patients were those with at least one administration of non-prophylactic antibiotics and who had at least one ICD-9 code indicating infection, while control patients were selected as those who had neither an ICD-9 code nor any administration of non-prophylactic antibiotics. Clinical suspicion of infection (and screening for the HAI cohort) was then derived using the administration time of first non-prophylactic antibiotics. We validated, in the MIMIC-III dataset, that the two criteria (microbiology culture test versus non-prophylactic antibiotics administration) yield a large overlap of the selected cohorts (see Supplementary material). Extraction of antibiotic records and non-prophylactic labelling details are also described in the Supplementary material.

Description of features and feature subsets used by the models

The extracted features are comprised of three sets of information: demographics (e.g., age, gender, height, weight), vital sign measurements (e.g., heart rate, blood pressure, temperature), and laboratory measurements (e.g., metabolic panels, complete blood count, and arterial blood gas). After feature extraction from each of the three hospital datasets, we applied an extensive preprocessing and cleaning pipeline to create a common and consistent dataset (see Supplementary material). A full list of the features is given in the Supplementary Table S1.

For training our machine learning algorithms, we defined an observation time as 1 h before each patient's clinical suspicion of infection (or randomly assigned event time for control patients). We then extracted the latest measured value of each feature leading up to the observation time and assembled these measurements into a physiological state vector for each patient. This feature vector was then augmented with features characterizing temporal trends from vital sign measurements during the 48 h window preceding the observation time, which was between 49 h before to 1 h before clinical suspicion (or randomly assigned event time for control patients). To mitigate sensitivity to outliers, we applied physiologic plausibility filters to the vital signs measured during the 48 h window before calculating trends. Trend features on laboratory measurements were excluded since they tend to be measured aperiodically (e.g., daily). Vital sign measurements, however, can have temporal resolution as high as every 5 min, e.g., in eICU dataset when data is consistently interfaced from bedside vital signs monitors into eCareManager. We extracted five trend features for the following vital signs: temperature, heart rate, systolic, diastolic, and mean blood pressures, oxygen saturation[1] (SpO_2), and respiration. For example, these trend features for heart rate are:

- Avg. (heart rate): the average heart rate value over a 48 h window.
- Min. (heart rate): the minimum heart rate value over a 48 h window.
- Max. (heart rate): the maximum heart rate value over a 48 h window.
- Var. (heart rate): the variance of heart rate over a 48 h window.
- CoefVar. (heart rate), or CV (heart rate): the coefficient of variation of heart rate over a 48 h window, defined as the standard deviation divided by the mean.

[1] Oxygen saturation is predominantly from pulse oximetry measurements and in addition blood gas measurements.

During the validation stage of our algorithm, we additionally applied the classifiers trained on the observation time of 1 h before clinical suspicion to earlier time windows in order to characterize predictive performance over time. These earlier observation times were 6 h, 12 h, 18 h, 24 h, and 48 h before clinical suspicion for infection patients (or randomly assigned event time for control patients). In those instances, we extracted a physiological state vector at earlier observation times in an analogous manner. For example, for the observation time of 6 h before clinical suspicion, we extracted the latest measured value of each feature leading up to 6 h before clinical suspicion and extracted trend features from vital sign measurements during the 48 h window preceding the observation time (that was between 54 h before to 6 h before clinical suspicion). Figure 2 provides a visual summary of the feature extraction pipeline.

Description of algorithms used

We employed two groups of algorithms: (a) linear classifiers, which identify a separating hyperplane in the original feature space; and (b) ensemble-based methods, which iteratively construct a powerful classifier from a set of "weak" nonlinear classifiers. We chose linear classifiers and ensemble-based methods over neural network techniques because we preferred to maintain interpretability of the trained model for clinical deployment, and to minimize the usage of computation resources to enable flexible applications. For linear classifiers we choose logistic regression, and for ensemble methods we benchmarked against abstained adaptive boosting with univariate decision stumps (20) and gradient boosting of decision trees using the XGBoost algorithm (21). Since our dataset is imbalanced in terms of infection prevalence, we employed stratified 5-fold cross-validation, and we did this for each of the three hospital datasets separately: with stratification, both the ratio of control to infection patients, and the ratio of patients from different hospital datasets are maintained in both training and testing sets. We compared model performance of different algorithms using the average model performance from the testing sets of the 5-fold cross-validation. Information about imputation, hyperparameter tuning and performance evaluation is detailed in the Supplementary material.

Description of model interpretation methods

The abstained adaptive boosting algorithm with decision stumps (20) can be expressed as a generalized additive model of the form $R(x) = r_1(x_1) + r_2(x_2) + \cdots + r_p(x_p)$ where $R(x)$ is the composite (ensemble) classifier, $x_1, x_2, \ldots, x_p$ are the p feature inputs, and $r_j(x_j)$, $j = 1, \ldots, p$ are the "weak learner" classifiers learned for each feature. In this case, infection patients are labeled as class 1 (controls are class −1), so that a larger value of $R(x)$ indicates the classifier's stronger confidence of the patient having infection. As a result, each $r_j(x_j)$ can be interpreted as an infection risk function evaluated with respect to a single feature. Because each $r_j(x_j)$ is the weighted sum of decision stumps acting on the respective feature, the infection risk of a single feature is a step function of the feature value, where each step is a decision threshold for different levels of infection risk. In order to control for the impact of feature missingness, we analyzed the relative importance of features through each $r_j(x_j)$ in two ways: (1) *total feature importance*, which evaluates a feature's importance across the entire cohort, and is calculated as the difference in the average infection risk between infection cohort and control cohort from the respective feature; and (2) *adjusted feature importance*, which isolates the feature's contribution on the subset of patients that have the feature measured, and is calculated as the difference in the average infection risk between infection cohort and control cohort that have the respective feature measured. Therefore, *total feature importance* gives an indication of a feature's effectiveness under typical hospital workflow conditions, while *adjusted feature importance* can identify discriminative features despite being less frequently measured.

The gradient boosting algorithm can be interpreted using SHAP (Shapley Additive exPlanations) method (22). SHAP assigns each feature an importance value for a particular prediction, therefore we can compare feature importance by examining the distribution of

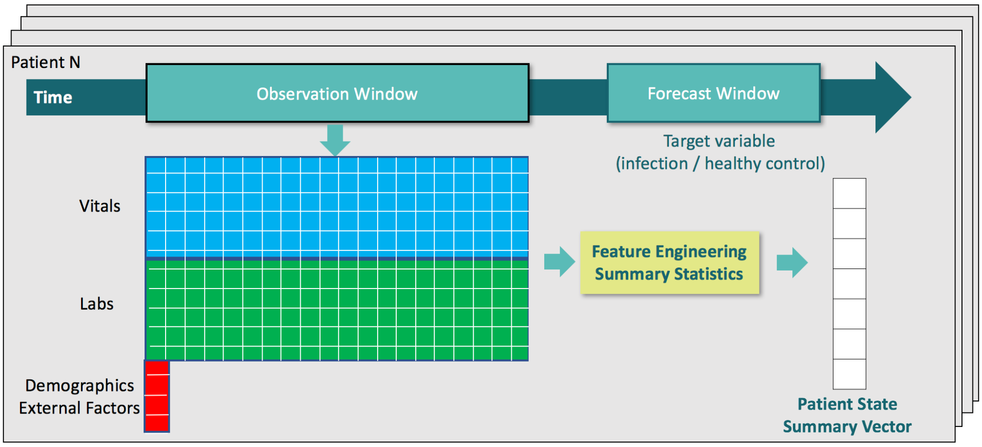

FIGURE 2
Diagram of the feature extraction pipeline.

SHAP values which represent the impacts each feature has on the model output.

Results

The cohort selection criteria resulted in a total training dataset size of 293,109 patients (256,327 control patients; 36,782 HAI patients). Of these patients, 63% are from the Banner Health dataset, 32% are from the eICU dataset, and 5% are from the MIMIC-III dataset. The majority of these patients are treated under ICU or general ward settings. Between the two infection cohort criteria (microbiology culture orders vs. non-prophylactic antibiotics administration), 26,599 HAI patients are identified from microbiology lab and ICD-9 code, while 10,183 infection patients are identified from non-prophylactic antibiotic administration and ICD-9 code.

Model performance

We compared machine learning algorithms in their ability to discriminate infection from control patients using clinical data acquired up to 1 h before clinical suspicion of infection. Our results show that gradient boosting with two level decision trees yielded the best performance with a mean AUC of 0.88 (standard deviation of 0.0009 from 5 testing folds), specificity of 0.93 and sensitivity of 0.54 at the break-even point (where sensitivity is approximately equal to positive predictive value (PPV), see Supplementary material), Sensitivity of 0.80 and 0.64, respectively, for when Specificity is 0.80 and 0.90 (Table 1: Xgboost). This performance was robust with different iterations of randomly down-sampled control cohort (AUC of 0.8839 ± 0.0003; mean ± standard deviation from 5 iterations). Abstained adaptive boosting with decision stump achieved a mean AUC of 0.85, specificity of 0.92 and sensitivity of 0.47 at break-even point, sensitivity of 0.73 and 0.54, respectively, for when specificity is 0.80 and 0.90 (Table 1: Abstained AdaBoost). Logistic regression performs poorly compared with ensemble algorithms, with a mean AUC of 0.77, specificity of 0.91 and sensitivity of 0.40 at break-even point, sensitivity of 0.60 and 0.43, respectively, for when specificity is 0.80 and 0.90 (Table 1: Logistic Regression). These results suggest that ensemble models are superior to linear models in predicting infection.

Next, we asked if ensemble models perform better than established empirical rules and clinical scores in infection prediction. First, fever or high body temperature (>98.6 F) is one of the first symptoms that lead to clinical suspicion of infection. Therefore, we compared temperature measurements between the infection and control cohorts and calculated the discriminative power of temperature at 1 h before clinical suspicion. Temperature by itself has an AUC = 0.59 for detecting infection, which is far inferior to performance achieved with gradient boosting (AUC = 0.88). Second, qSOFA—quick sequential organ failure assessment—was introduced by the Third International Consensus Definitions for Sepsis and Septic Shock task force in 2016, and is proposed as a quick assessment tool for identifying sepsis among patients with infection (23). Based on the Sepsis-3 criteria, we extracted Glasgow Coma Score, Systolic Blood Pressure, and Respiratory Rate from the medical database, and derived qSOFA scores at 1 h before clinical suspicion of infection. In total 111,651 qSOFA scores were extracted, 22,460 from infection cohort and 89,191 from control cohort (infection prevalence = 20.1%). We then calculated the area under ROC curve of infection prediction by using qSOFA alone. qSOFA by itself has an AUC = 0.59 when predicting infection at 1 h before suspicion of infection. To ensure a fair comparison with ensemble models, we re-trained the gradient boosting algorithm using data from the subset of patient cohort that have qSOFA available. Gradient boosting on the patient subset achieves an AUC of 0.83 which is substantially better than the performance of qSOFA. Overall our results suggest advantages of ensemble models over established clinical methods in infection prediction.

We further benchmarked ensemble model performance when feature sets are reduced. First, we excluded all lab measurements and focused on 14 vital signs and demographics factors (plus 50 derived trend features), as they are continuously available and more predictably available than lab measurements. Gradient boosting, re-trained from the feature space excluding labs, achieved a mean AUC of 0.81, specificity of 0.92 and sensitivity of 0.42 at break-even point, sensitivity of 0.62 and 0.45, respectively, for when specificity is 0.80 and 0.90 at 1 h before clinical suspicion of infection (Table 1: GradientBoost—exclude lab). Second, we excluded infrequently measured features that are available for less than 70% of the patient cohort. This produced a reduced feature space with 36 vitals, demographics and laboratory measurements (plus 32 derived trend features). Gradient boosting model, re-trained from frequently measured features, achieved a mean AUC of 0.86, specificity of 0.93 and sensitivity of 0.50 at break-even point, sensitivity of 0.74 and 0.57, respectively, for when specificity is 0.80 and 0.90 at 1 h before clinical suspicion of infection (Table 1: Xgboost—reduced features). These results suggest that it is possible to obtain good performance when reducing the total feature space by half.

In addition, we investigated the infection prediction performance of ensemble models at earlier time points. We applied the most interpretable model (Abstained AdaBoost) and the best performing model (Gradient

TABLE 1 Performance of infection prediction at 1 h before clinical suspicion of infection.

Algorithm	AUC	Sensitivity (spec) break-even point	Sensitivity @ specificity = 0.8	Sensitivity @ specificity = 0.9
GradientBoost	0.884	0.537 (0.934)	0.800	0.635
Abstained AdaBoost	0.852	0.469 (0.924)	0.731	0.536
Logistic Regression	0.772	0.399 (0.914)	0.597	0.431
GradientBoost—exclude lab	0.810	0.415 (0.916)	0.622	0.449
GradientBoost—reduced features	0.862	0.499 (0.928)	0.750	0.574

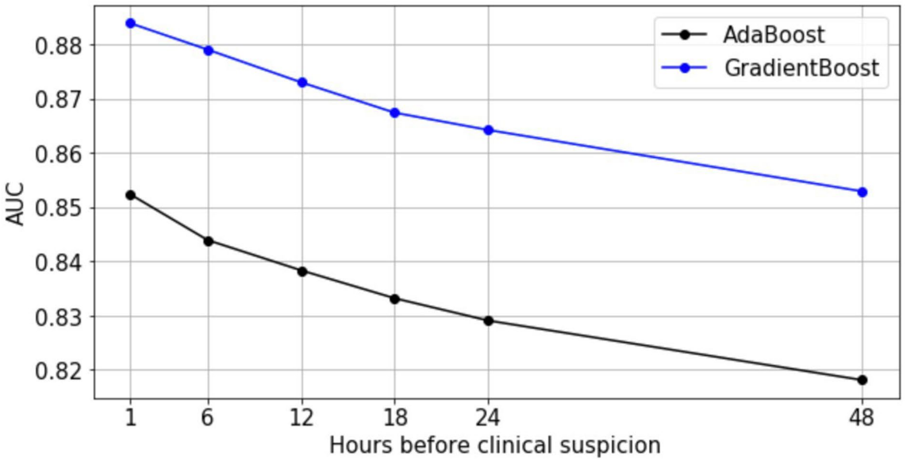

FIGURE 3
Predictive performance of AdaBoost and GradientBoost models relative to time of clinical suspicion.

TABLE 2 Feature importance rankings from abstained AdaBoost model (top 15).

Total feature importance		Adjusted feature importance	
Rank	Feature	Rank	Feature
1	Albumin	1	Albumin
2	Max. (SpO_2)	2	TIBC
3	pH	3	Fibrinogen
4	Min. (SpO_2)	4	Temperature
5	Temperature	5	ESR
6	Avg. (SpO_2)	6	PVRI
7	Var. (SpO_2)	7	Max. (Temperature)
8	Lactate	8	Urinary RBC
9	Bands	9	Avg. (Respiration)
10	Max. (Temperature)	10	WBC
11	Avg. (Respiration)	11	BUN
12	CV (SpO_2)	12	CRP
13	FiO_2	13	Ferritin
14	WBC	14	Neutrophils
15	Bicarbonate	15	Var. (Temperature)

Total feature importance evaluates a feature's importance across the entire cohort; adjusted feature importance isolates the feature's contribution on the subset of patients that have the feature measured.

Boosting) to earlier observation windows to characterize predictive performance over time using the full feature space (Figure 3). Despite degraded model performance over time, gradient boosting maintains an AUC >0.85, while abstained adaptive boosting maintains an AUC >0.81 for 48 h before clinical suspicion. These results support an assertion that it is possible to predict hospital acquired infection earlier, up to 48 h before clinical suspicion of infection.

Model interpretation

To better understand the biomarkers leveraged by the ensemble-based models, we first analyze the AdaBoost algorithm with decision stumps since it is easier to interpret, and then contrast with feature importance scores on the GradientBoost algorithm with decision trees using the SHAP (Shapley additive explanations) method (22).

We first examined the top 15 features ranked by *total feature importance* and *adjusted feature importance* derived from abstained adaptive boosting model trained in the full feature space (Table 2). As described in Methods, *total feature importance* evaluates a feature's importance across the entire cohort, and *adjusted feature importance* isolates the feature's contribution on the subset of patients that have the feature measured. From both metrics, we found that the top 15 features are a mix of laboratory measurements and vital signs. Adjusted feature importance, in particular, identifies discriminative features from laboratory measurements despite being less frequently measured.

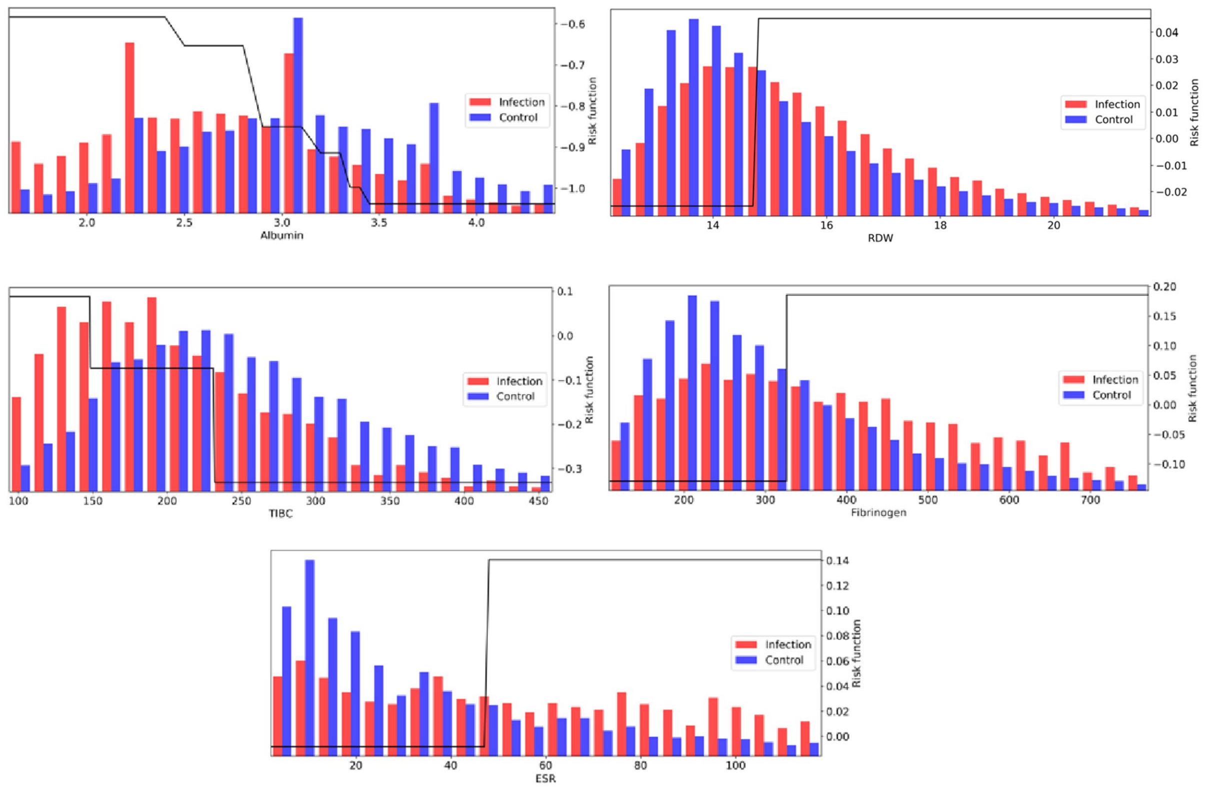

FIGURE 4
AdaBoost risk functions (black) for a subset of the most important laboratory measurements, along with population distribution underlays for infection (red) and control (blue) populations.

The learned risk functions behave in clinically interpretable ways. Figure 4 visualizes the risk functions (black) for a subset of the most important laboratory features, along with population distribution underlays for infection (red) and control (blue) populations. The learned risk functions for these representative features are either monotonically increasing, suggesting that an elevation of the respective clinical measurement is associated with higher infection risk; or monotonically decreasing, suggesting that a decrease of the respective clinical measurement is associated with higher infection risk. During training, each risk function is assembled from a collection of decision stumps that identify key feature thresholds that distinguish levels of infection risk. The scale of the risk function (the y-axis in Figure 4 plots) is unitless, but can be used to compare the relative importance of features (see Methods and Table 2 for further details on feature importance).

Amongst laboratory measurements, a number of features associated with, but not necessarily specific to, inflammation were identified. The top feature across both scoring metrics was associated with hypoalbuminemia (low albumin levels <3 g/dL), which has been shown to correlate with inflammation, shock, and sepsis (24). High RDW (>15%) was also a strong biomarker, with literature showing it correlated with inflammation markers CRP and ESR (14). With respect to the *adjusted feature importance* score, a number of infrequently measured features, but highly discriminative, were identified by the model, all of which show associations with inflammatory response: low TIBC (<240 mcg/dL; prevalence = 3%), elevated Fibrinogen (>325 mg/dL; prevalence = 5%), and elevated ESR (>45 mm/h; prevalence = 2%).

Many other laboratory values were also discriminative. Increased risk is identified when Bicarbonate levels fall below approximately 24 mEq/L, which may be indicative of metabolic acidosis, in particular lactic acidosis (elevated Lactate levels above 1.5 mmol/L were also contributing to infection risk). White blood cell concentrations (25) were also strong indicators in the top 15 features, with elevated bands and neutrophil concentrations. Other notable indicators are low HDL and LDL cholesterol levels (26), and increases in blood platelets, which is a sign of host defense and induction of inflammation and tissue repair in response to infection onset (27).

Although laboratory measurements play a significant role, the model also aggregates information from a number of vital signs. The infection risk function based on temperature increases rapidly above 37.8°C, although this accounts for a small percentage of infection patients (5,105 out of 40,406 (~12.6%) of infection patients registered a fever ≥37.8°C at the 1 h window). For controls, 5,579 out of the 96,505 control patients (~5.8%) exhibited a fever ≥37.8C. Infection patients tend to have an elevated heart rate and macro variability, which is reported to be critical for the diagnosis and prognosis of infection by many studies (28, 29). For blood pressure, patients tend to have a decreased blood pressure (systolic, diastolic, and mean), and this effect was often selected by the classifier. Many trend variability features on vitals were selected across temperature, heart rate, blood pressure, oxygen saturation (SpO_2), and respiration, as the infection cohort tends to exhibit a heavier right tail in feature variance measures.

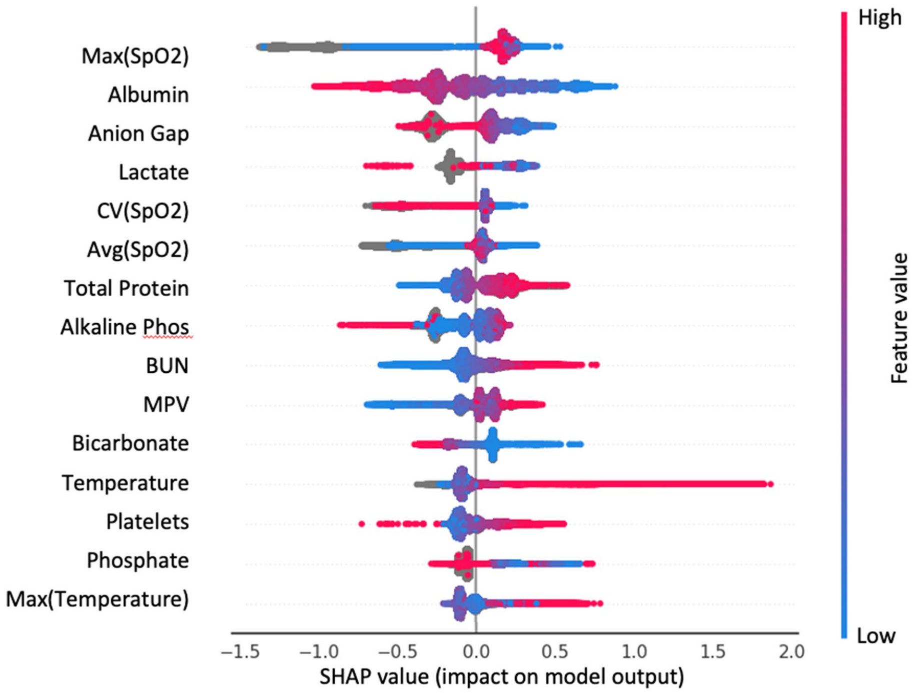

FIGURE 5
Top 15 important features of the GradientBoost model from SHAP analysis. Each dot is a patient; color indicates the value of the feature. SHAP value is on the *x*-axis: large positive value—feature contributes strongly to predict infection; large negative value—feature contributes strongly to predict control.

Changes in vital signs are also reported in the literature to accompany the development of infection (30, 31).

We additionally applied SHAP analysis to extract feature importance rankings from the gradient boosting method (Figure 5). We have observed overlaps in the selected features between the more interpretable AdaBoost model and gradient boosting, such as albumin, SpO_2, bicarbonate, temperature, lactate and BUN.

Algorithm performance on infection subgroups

Patients' host responses to pathogens vary between pathogens and primary sites of infection which result in heterogeneous physiological changes. The extracted HAI cohort is mainly from, ranked by high to low prevalence, the following five infection types (defined by ICD-9 codes—see Supplementary material): pneumonia (17,224 patients), bloodstream infection (12,891 patients), bone/joint/tissue/soft tissue infection (11,613 patients), sepsis (9,643 patients) and urinary system infection (9,118 patients). Note that these patients are primarily from ICUs or general wards, and some patients can have more than one HAI. To compare detection performance on different infection types, we calculated recall (Sensitivity) from the model for patient subgroups of different infection types (Figure 6). We found that the infection model (Table 1: Xgboost) has the highest recall in predicting Sepsis (recall = 0.70) and bloodstream infection (recall = 0.67), followed by pneumonia (recall = 0.61), bone/joint/tissue/soft tissue infection (recall = 0.50) and urinary system infection (recall = 0.46). This result indicates that the infection model performs the best in predicting subgroups of patients that have high acuity.

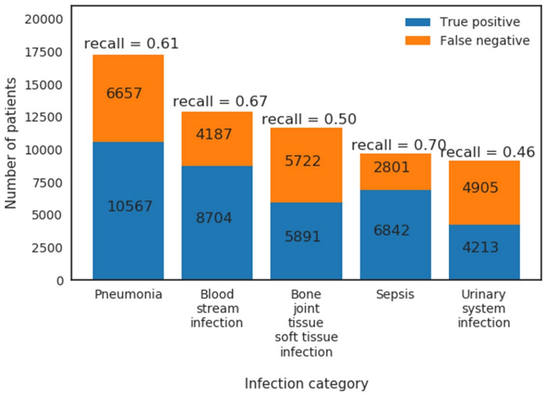

FIGURE 6
True positives and false negatives from GradientBoost model for the top five prevalence infection categories.

Impact of comorbidities on algorithm performance

The previous section assessed true positive rates (recall/sensitivity) for various infection types. By the same token, we may also characterize true negative performance of the algorithm with respect

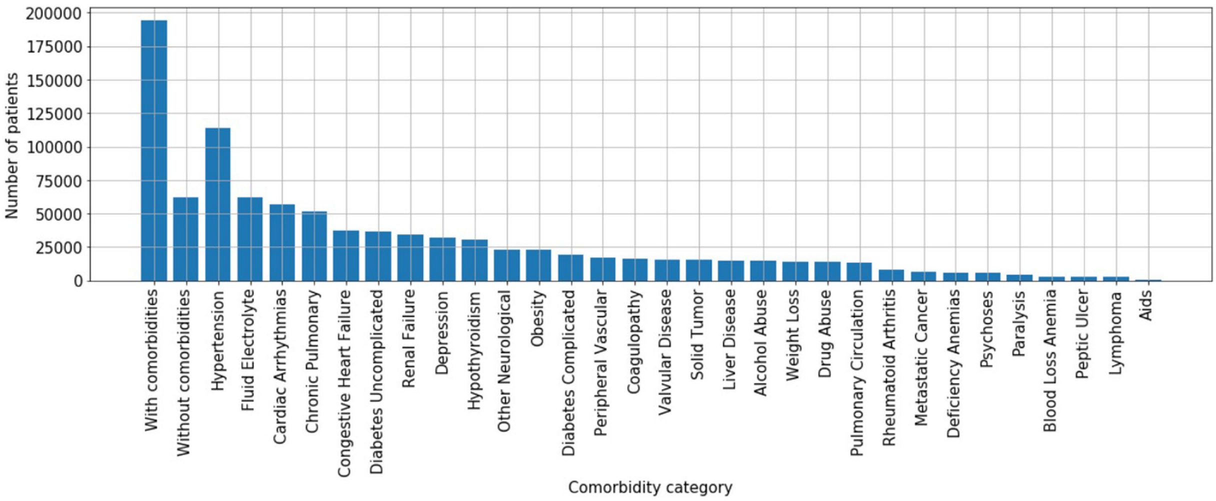

FIGURE 7
Comorbidity prevalence amongst control patients.

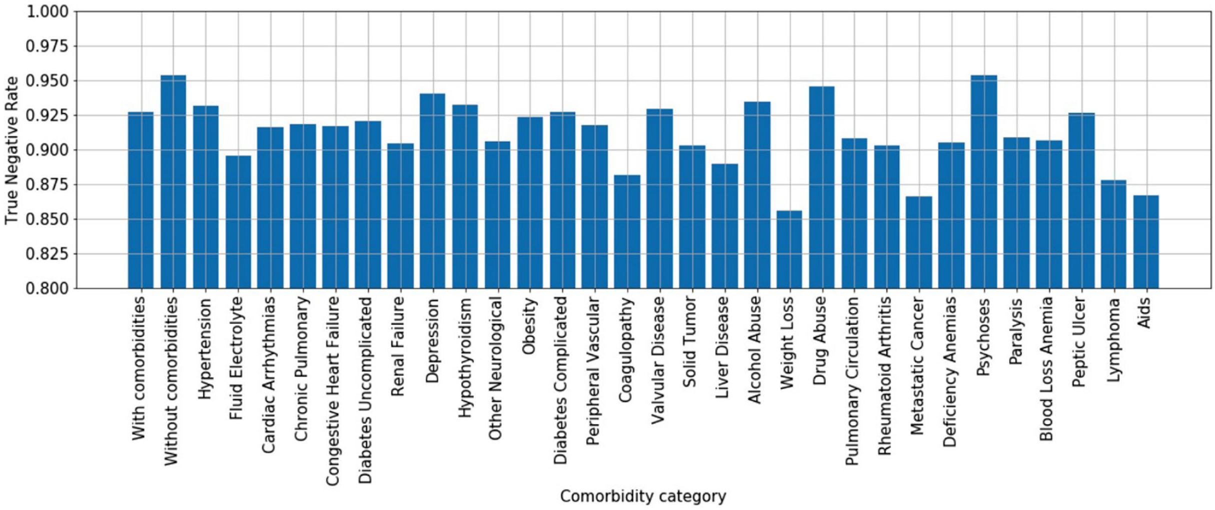

FIGURE 8
True negative rates (specificity) by comorbidity category. *x*-axis: "with comorbidities"—control patients with at least one comorbidity; "without comorbidities"—control patients without any documented comorbidities; 30 comorbidity categories are ordered by prevalence shown in Figure 6 to highlight that the differences in true negative rate are not simple reflections of prevalence.

to various chronic comorbidities exhibited by the control patient population. To do so, we calculated the Elixhauser Comorbidity index (32) for each control patient, which associates diagnostic ICD-9 codes [see Table 2 of (32)] with a set of 30 comorbidity categories. Of the 256,327 control patients, 194,364 (76%) exhibited at least one comorbidity—see Figure 7 for a summary of prevalence of each comorbidity category amongst control patients. We then calculated the infection model's true negative rate (TNR) on the control patient population that exhibited each of the 30 comorbidity categories. In addition, we compared true negative rate for control patients with at least one comorbidity (76% of all control patients, labeled "with comorbidities") to the true negative rate for control patients without any documented comorbidities (24% of all control patients, labeled "without comorbidities")—see Figure 8.

The model performs better at ruling out infection on control patients without comorbidities than those with comorbidities (TNR = 0.95 vs. TNR = 0.925), suggesting that confounding chronic conditions contribute to the false positive rate of the model. Interestingly, with respect to individual comorbidity categories, the model performs best at ruling out infection on control patients with neurological comorbidities (e.g., depression, psychoses), drug/alcohol abuse, and hypothyroidism; presumably since such conditions may have limited overlap in physiological biomarkers related to infection. The worst performing comorbidity categories include fluid/electrolyte disorders, coagulopathy, weight loss, metastatic cancer, lymphoma, anemia, and AIDS.

Discussion

Our work addresses the fundamental problem of early prediction of HAI, to allow prompt treatment and prevention of infectious disease transmission. We presented a large-scale, retrospective big data machine learning study that provides a data-driven approach to the problem, which can be tailored and adapted to different populations of interest. Infection can be detected by our model with high accuracy in its pre-symptomatic state at 48 h before clinical suspicion.

The training data of 293,109 patients for our infection prediction model was curated from three large hospital datasets that included patient encounters under both high-acuity and low-acuity settings from >400 US hospitals in the span of 16 years. The purpose of using such a large scale dataset for training was to enable the infection prediction model to learn from a heterogeneous patient cohort and to accommodate different data availability and frequencies under different care settings. An extensive preprocessing and data cleaning pipeline was developed to create a common and consistent dataset across the hospitals and acuity settings (see Supplementary material). Because the model was not biased by a single hospital or a single dataset, it should generalize well in real-world use cases in predicting infection.

Ensemble models proved to perform significantly better than both the established empirical rules and clinical scores, and logistic regression, with gradient boosting having the best performance. AdaBoost provided an interpretable model which allows us to map the feature importance to its relevance in clinical literature. For example, multiple laboratory values associated with inflammation ranked high in the feature importance metric, as well as features indicative of acidosis. High heart rate, high temperature and macro variability of vital signs were also indicative of infection, consistently with what has been reported in the literature (28–31). This characteristic of interpretability not only further validates our model, but also provides meaningful information in the clinical setting, quantifying the effect that appropriate action on each of these parameters would have in preventing HAI. It is well known that interpretability of the decision support model is vital to the acceptance of such a predictor in the clinical setting (33).

One important finding of our study is that the high performance of the model is obtained only by aggregating multiple biomarkers. No single "super feature" exists that allows superior classification. This likely reflects at the same time the variable etiology of the HAI—which can be of different natures (respiratory, blood stream infection, sepsis, etc.), the individual variability in the response, and the multi-system nature of the effect of the infection on the patient's physiology. On the other hand, it is still possible to obtain prediction performance that are clinically viable with a reasonable number of clinical measurements. We have showed that with a core set of 36 clinical measurements, the infection model performs at an AUC = 0.86 at 1 h before clinical suspicion of infection.

The algorithm presented in this work could be implemented in a hospital setting by leveraging the existing monitoring systems and infrastructure. When risk of infection is predicted in advance, knowledge of the contributing parameters provided by the transparency of the model would allow secondary assessment and prompt intervention. While the best performing model employs a combination of laboratory test values and vital signs across 163 features, a model trained on 36 of the most frequently measured vital signs, labs and demographics achieves an AUC of 0.86 at 1 h before clinical suspicion. Moreover, a model trained with only vital signs and demographics still achieves an acceptable area under the curve, equal to 0.81. A similar model could be employed in a context that is outside of the hospital (e.g., home monitoring via wearable devices) or in other situations where laboratory values are not easily obtainable.

Limitations

In this section we describe a couple of limitations on our study due to the complex nature of analyzing large retrospective hospital datasets.

First, we tested our model using six different observation times that were 1 h, 6 h, 12 h,18 h, 24 h, and 48 h before clinical suspicion of infection. This design warranted us to have at least an hour of prediction gap before the labeled time of clinical suspicion of infection. This was because determining the exact timing of clinical suspicion of infection was difficult and might not be possible, a prediction gap was built into account for the time differences between the true clinical suspicion of infection and when the culture was ordered in the EMR system. We reasoned in high-acuity settings such as ICUs this 1 h gap was sufficient. For the general ward encounters in Banner dataset, clinical suspicion of infection may arise a couple of hours before the ordering of microbiology culture test given the typical workflow in that environment. In this case it is more accurate to look at the performance at the observation time of 6 h before clinical suspicion instead of 1 h to evaluate the model in predicting infection shortly before the true clinical suspicion of infection (we reported AUC = 0.88 at 6 h before clinical suspicion, Figure 3 blue line).

Second, our infection and control cohort selection criteria were designed to be conservative, in that we only included patients in infection cohort if they satisfied both criteria (ICD-9 and microbiology) and only included patients in control cohort if they met none of the two criteria. This means that we excluded, from the infection cohort, those patients who had an infection diagnosis but whose timing of clinical suspicion of infection could not be localized; and that we excluded, from the control cohort, those patients who had a microbiology culture test ordered but did not have an infection diagnosis. For the latter patient group, some of them may have a negative culture but the culture was ordered based on clinical suspicion. It would be interesting to examine the model performance in those patients. We suspect, because those patients may have overlap in symptomatology (hence the clinical suspicion) and physiological biomarkers related to infection, our model may have a degraded performance in true negative rates in this group of patients.

Finally, the patient encounters used in this study happened before the full adoption of ICD-10 therefore we used ICD-9 to select the infection patients. We understand that ICD-10 have improved granularity over ICD-9 therefore are more specific in identifying health conditions. For training a general infection prediction model where different types of infections were grouped in the same category, we believe the granularity provided in ICD-9 is sufficient. However, it would be interesting to see how using ICD-10 would affect the model performance in different infection categories (Figure 6).

Conclusion

This study developed an algorithm for early identification of infection in hospitalized patients, using machine learning applied to large retrospective hospital datasets. The model is able to identify patients who are infected with reasonable performance up to 48 h before clinical suspicion of infection (AUC >0.85). The trained models utilize ensembles of decision trees, which are readily interpretable and provide ranked lists of feature importance. The primary model leveraging all available (163) vital signs, laboratory measurements and demographics achieves the best performance; however, a secondary model limited to the 36 most commonly measured clinical measurements still achieves an AUC = 0.86 at 1 h before clinical suspicion. The models compare favorably to established clinical rules and show high potential for real-world hospital deployment as a clinical decision support aid.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: MIMIC-III dataset is available in PhysioNet repository, https://mimic.physionet.org/. A portion of the eICU dataset used in this study is available in PhysioNet repository, https://eicu-crd.mit.edu; the remaining of the eICU dataset is proprietary to Philips. The Banner Health dataset is a proprietary dataset that is not publicly shareable. Requests to access these datasets should be directed to Banner Health and Philips.

Ethics statement

The studies involving humans were approved by the following ethics committee/institutional review board: The MIMIC-III project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Use of the eICU data was approved by the Philips Internal Committee for Biomedical Experiments. Banner Health data use was a part of an ongoing retrospective deterioration detection study approved by the Institutional Review Board of Banner Health and by the Philips Internal Committee for Biomedical Experiments. Requirement for individual patient consent was waived because the project did not impact clinical care, was no greater than minimal risk, and all protected health information was removed from the limited dataset used in this study. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

TF, CK, and BC participated in the study design, data preparation and analysis, machine learning model training, and contributed to writing of the manuscript. DN, SM, CZ, EG, and DM contributed to study design, data analysis, and contributed to writing of the manuscript. DS provided clinical consultation, manuscript review, and interpretation of results. JF provided clinical consultation and participated in hypothesis development, cohort identification, and manuscript review and interpretation of results. All authors contributed to the article and approved the submitted version.

Funding

This work is sponsored by the US Department of Defense (DoD), Defense Threat Reduction Agency (DTRA) under project CB10560. The funding body did not play a role in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Conflict of interest

TF, DN, CK, SM, EG, and BC are employees of Philips Research. CZ, JF, and DM were employees of Philips Research. DS is employee of Banner Health.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1213411/full#supplementary-material

References

1. Centers for Disease Control and Prevention, *2018 national and state healthcare-associated infections progress report*, (2019). Available at: http://www.cdc.gov/hai/data/portal/progress-report.html

2. Klevens RM, Edwards JR, Richards CL Jr, Horan TC, Gaynes RP, Pollock DA, et al. Estimating health care-associated infections and deaths in U.S. hospitals, 2002. *Public Health Rep*. (2007) 122:160–6. doi: 10.1177/003335490712200205

3. Baker MA, Sands KE, Huang SS, Kleinman K, Septimus EJ, Varma N, et al. The impact of coronavirus disease 2019 (COVID-19) on healthcare-associated infections. *Clin Infect Dis*. (2022) 74:1748–54. doi: 10.1093/cid/ciab688

4. RD MA, Miller M, Albertson T, Panacek E, Johnson D, Teoh L, et al. Adequacy of early empiric antibiotic treatment and survival in severe sepsis: experience from the MONARCS trial. *Clin Infect Dis*. (2004) 38:284–8. doi: 10.1086/379825

5. Ferrer R, Artigas A, Suarez D, Palencia E, Levy MM, Arenzana A, et al. Effectiveness of treatments for severe sepsis. *Am J Respir Crit Care Med*. (2009) 180:861–6. doi: 10.1164/rccm.200812-1912OC

6. Longini IM Jr, Halloran ME, Nizam A, Yang Y. Containing pandemic influenza with antiviral agents. *Am J Epidemiol*. (2004) 159:623–33. doi: 10.1093/aje/kwh092

7. Pronovost PJ, Goeschel CA, Wachter RM. The wisdom and justice of not paying for "preventable complications". *JAMA*. (2008) 299:2197–9. doi: 10.1001/jama.299.18.2197

8. Wilke WS, Long JK, Mossad SB, Goldman MP. Antiviral agents for treating influenza. *Cleve Clin J Med*. (2000) 67:92–5. doi: 10.3949/ccjm.67.2.92

9. Churpek MM, Snyder A, Sokol S, Pettit NN, Edelson DP. Investigating the impact of different suspicion of infection criteria on the accuracy of quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores. *Crit Care Med*. (2017) 45:1805–12. doi: 10.1097/CCM.0000000000002648

10. Bhattacharjee P, Edelson DP, Churpek MM. Identifying patients with sepsis on the hospital wards. *Chest*. (2017) 151:898–907. doi: 10.1016/j.chest.2016.06.020

11. Umscheid CA, Betesh J, VanZandbergen C, Hanish A, Tait G, Mikkelsen ME, et al. Development, implementation, and impact of an autoated early warning and response system for sepsis. *J Hosp Med*. (2015) 10:26–31. doi: 10.1002/jhm.2259

12. Holmes J, Roberts G, Meran S, Williams JD, Phillips AOWelsh AKI Steering Group. Understanding electronic AKI alerts. *Kidney Int Rep*. (2017) 2:342–9. doi: 10.1016/j.ekir.2016.12.001

13. Holmes J, Rainer T, Geen J, Roberts G, May K, Wilson N, et al. Acute kidney injury in the era of the AKI E-alert. *Clin J Am Soc Nephrol*. (2016) 11:2123–31. doi: 10.2215/CJN.05170516

14. Jones MN. NEWSDIG: The National Early Warning Score Development and Implementation Group. *Clin Med*. (2012) 12:501–3. doi: 10.7861/clinmedicine.12-6-501

15. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual*. (2017) 6:e000158. doi: 10.1136/bmjoq-2017-000158

16. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*. (2017) 6:e000158. doi: 10.1136/bmjresp-2017-000234

17. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One*. (2017) 12:e0174708. doi: 10.1371/journal.pone.0174708

18. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016) 3:1–9. doi: 10.1038/sdata.2016.35

19. Pollard TJ, AEW J, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data*. (2018) 5:180178. doi: 10.1038/sdata.2018.178

20. Conroy B, Eshelman L, Potes C, Xu-Wilson M. A dynamic ensemble approach to robust classification in the presence of missing data. *Mach Learn*. (2016) 102:443–63. doi: 10.1007/s10994-015-5530-z

21. Chen T., Guestrin C., "Xgboost: a scalable tree boosting system, "22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016).

22. Lundberg S. M., Lee S.-I., A unified approach to interpreting model predictions, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, (2017).

23. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. (2016) 315:801–10. doi: 10.1001/jama.2016.0287

24. Nicholson JP, Wolmarans MR, Park GR. The role of albumin in critical illness. *Br J Anaesth*. (2000) 85:599–610. doi: 10.1093/bja/85.4.599

25. High white blood cell count. *Mayo Clinic*. Available at: https://www.mayoclinic.org/symptoms/high-white-blood-cell-count/basics/definition/sym-20050611

26. Pirillo A, Catapano AL, Norata GD. HDL in infectious diseases and sepsis. *Handb Exp Pharmacol*. (2015) 224:483–508. doi: 10.1007/978-3-319-09665-0_15

27. Klinger MHF, Jelkmann W. Review: role of blood platelets in infection and inflammation. *J Interf Cytokine Res*. (2002) 22:913–22. doi: 10.1089/10799900260286623

28. Ahmad S, Tejuja A, Newman KD, Zarychanski R, Seely AJ. Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection. *Crit Care*. (2009) 13:232–2. doi: 10.1186/cc8132

29. Karmali SN, Sciusco A, May SM, Ackland GL. Heart rate variability in critical care medicine: a systematic review. *Intensive Care Med Exp*. (2017) 5:33–3. doi: 10.1186/s40635-017-0146-1

30. González Plaza JJ, Hulak N, Zhumadilov Z, Akilzhanova A. Fever as an important resource for infectious diseases research. *Intractable Rare Dis Res*. (2016) 5:97–102. doi: 10.5582/irdr.2016.01009

31. Hamano J, Tokuda Y. Changes in vital signs as predictors of bacterial infection in home care: a multi-center prospective cohort study. *Postgrad Med*. (2017) 129:283–7. doi: 10.1080/00325481.2017.1251819

32. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. (2005) 43:1130–9. doi: 10.1097/01.mlr.0000182534.19832.83

33. Huang Q, Sun Y, Jia M, Zhang T, Chen F, Jiang M, et al. Quantitative analysis of the effectiveness of antigen- and polymerase chain reaction-based combination strategies for containing COVID-19 transmission in a simulated community. *Engineering*. (2023). doi: 10.1016/j.eng.2023.01.004