Machine Learning-based Clinical Decision Support for Infection Risk Prediction

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

[1] Philips Research North America, Cambridge MA, USA

[2] Philips Research Bangalore, Bengaluru, India

[3] Department of Medical Informatics, Banner Health, Phoenix AZ, USA

[4] Department of Biomedical Informatics, University of Arizona College of Medicine, Phoenix AZ, USA

[5] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge MA, USA

* Corresponding author

Corresponding author email: bryan.conroy@philips.com

ABSTRACT

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, which is defined as the order of a microbiology test or administration of antibiotics.

Results: We find that our best performing infection risk model achieves a cross-validated AUC of 0.88 at 1-hour before clinical suspicion and maintains an AUC>0.85 for 48-hours 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-hour 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 on model interpretability via feature importance rankings.

Conclusions: 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)

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 cost 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] [10] [11]. Many of these CDS tools are

60 rule based and developed through physician consensus and guidelines. These include more
61 standardized solutions like the Acute Kidney Injury (AKI) eAlert that has been deployed in
62 hospitals in Wales [12] [13] and the National Early Warning Score (NEWS) score that is
63 standard for detecting general clinical deterioration in the UK [14]. While these approaches
64 benefit from clinician experience, they are simplified to remain generalizable and fail to capture
65 the complete clinical context required to discriminate difficult or atypical cases. In addition,
66 these approaches are not easily tailored or adapted, for example, to specific patient populations.
67 More recently, several studies have suggested data-driven approaches to create physiological risk
68 prediction algorithms, including in the areas of infection and sepsis prediction [9] [15] [16] [17].

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

79 **METHODS**

80 ***Description of data***

81 We combined clinical data from three large hospital datasets: the MIMIC-III (Medical
82 Information Mart for Intensive Care III) database collected from 2001 to 2012 [18], the eICU
83 dataset from Philips' electronic ICU telemedicine business collected from 2003 to 2016 [19], and
84 a dataset of electronic medical records from Banner Health collected from 2010 to 2015. In total,
85 the combined dataset includes over 6.5 million patient encounters collected from more than 450
86 hospitals. Supplemental Figure 1 indicates the types of data present in each hospital dataset.

87 ***Ethical Approval***

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

97 ***Infection and control cohort extraction***

98 We define infection patients as those who 1) have a confirmed infection diagnosis, and 2) have
99 data indicating clinical suspicion of infection. Patients in the infection cohort were selected as
100 those with confirmed infection diagnoses via ICD-9 and whose timing of clinical suspicion of

101 infection could be localized by a microbiology culture test order. Infection patients were then
102 further screened into an HAI cohort if the timing of clinical suspicion of infection occurred at
103 least 48 hours after admission. Patients in the control cohort were selected as those who have
104 neither an infection-related ICD-9 diagnosis code nor any microbiology culture tests ordered.
105 Since the selection criteria identified a much larger set of control patients than HAI patients, we
106 down-sampled the control cohort population to maintain a prior infection odds (prevalence) of
107 12.5%. This ensured that the training dataset would not be overly dominated by control patients,
108 while still maintaining the HAI cohort as the minority class. Figure 1 shows the general
109 decision scheme behind this methodology. Curation of infection ICD-9 codes is described in
110 detail in the Supplementary Materials.

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

For control patients, we generated a synthetic event time, such that clinical data used for prediction 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 hours after the patient's first clinical measurement, and that precedes the end of the patient's hospital stay as the synthetic event time.

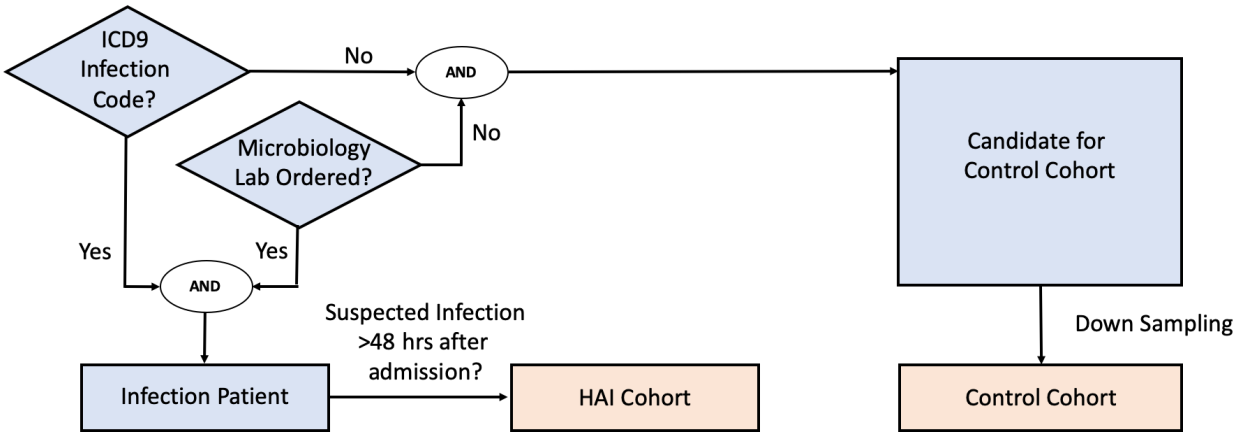


Figure 1: Cohort inclusion/exclusion criteria flow diagram

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. A full list of the features is given in the Supplemental Materials in Table A1.

136 For the purpose of training our machine learning algorithms, we defined an observation time as
137 one hour before each patient's clinical suspicion of infection (or randomly assigned event time
138 for control patients). We then extracted the latest measured value of each feature leading up to
139 the observation time and assembled these measurements into a physiological state vector for
140 each patient. This feature vector was then augmented with features characterizing temporal
141 trends from vital sign measurements during the 48-hour window preceding the observation time.
142 To mitigate sensitivity to outliers, we applied physiologic plausibility filters to the vital signs
143 before calculating trends. Trend features on laboratory measurements were excluded since they
144 tend to be measured aperiodically (e.g., daily). We extracted five trend features for the for vital
145 signs: Temperature, Heart Rate, Systolic, Diastolic, and Mean Blood Pressures, Oxygen
146 Saturation[1] (SpO2), and Respiration. For example, these trend features for Heart Rate are:
147 Avg(Heart Rate): The average heart rate value over a 48-hour window

148 - Min(Heart Rate): The minimum heart rate value over a 48-hour window
149 - Max(Heart Rate): The maximum heart rate value over a 48-hour window.
150 - Var(Heart Rate): The variance of heart rate over a 48-hour window
151 - CoefVar(Heart Rate), or CV(Heart Rate): The coefficient of variation of heart rate over a
152 48-hour window, defined as the standard deviation divided by the mean

153 During the validation stage of our algorithm, we additionally applied the classifiers trained on
154 the one-hour before observation time to earlier time windows in order to characterize predictive
155 performance over time. In those instances, we extracted a physiological state vector at earlier

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

156 observation times in an analogous manner. Figure S3 provides a visual summary of the feature
157 extraction pipeline.

158 ***Description of algorithms used***

159 We employed two groups of algorithms: (a) linear classifiers, which identify a separating
160 hyperplane in the original feature space; and (b) ensemble-based methods, which iteratively
161 construct a powerful classifier from a set of "weak" nonlinear classifiers. We chose linear
162 classifiers and ensemble-based methods over neural network techniques because we preferred to
163 maintain interpretability of the trained model for clinical deployment, and to minimize the usage
164 of computation resources to enable flexible applications. For linear classifiers we choose logistic
165 regression, and for ensemble methods we benchmarked against Abstained Adaptive Boosting
166 with univariate decision stumps [20] and Gradient Boosting of decision trees using the XGBoost
167 algorithm [21]. Since our dataset is imbalanced in terms of infection prevalence, we employed
168 stratified cross-validation, and we did this for each of the three hospital datasets separately: with
169 stratification, both the ratio of control to infection patients, and the ratio of patients from
170 different hospital datasets are maintained in both training and testing sets. Information about
171 imputation, hyperparameter tuning and performance evaluation is detailed in the Supplemental
172 Materials.

173 ***Description of model interpretation methods***

174 The Adaptive Boosting algorithm with decision stumps can be expressed as a generalized
175 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
176 (ensemble) classifier, $x_1, x_2, \ldots, x_p$ are the p feature inputs, and $r_j(x_j)$, j=1,…,p are the "weak

177 learner" classifiers learned for each feature. In this case, infection patients are labeled as class 1
178 (controls are class -1), so that a larger value of R(x) indicates the classifier's stronger confidence
179 of the patient having infection. As a result, each $r_j(x_j)$ can be interpreted as an infection risk
180 function evaluated with respect to a single feature. In order to control for the impact of feature
181 missingness, we analyzed the relative importance of features through each $r_j(x_j)$ in two ways: (1)
182 *total feature importance*, which evaluates a feature's importance across the entire cohort; and (2)
183 *adjusted feature importance*, which isolates the feature's contribution on the subset of patients
184 that have the feature measured. Therefore, *total feature importance* gives an indication of a
185 feature's effectiveness under typical hospital workflow conditions, while *adjusted feature*
186 *importance* can identify discriminative features despite being less frequently measured.

187 The Gradient Boosting algorithm can be interpreted using SHAP (Shapley Additive
188 exPlanations) method [22]. SHAP assigns each feature an importance value for a particular
189 prediction, therefore we can compare feature importance by examining the distribution of SHAP
190 values which represent the impacts each feature has on the model output.

191 **RESULTS**

192 The cohort selection criteria resulted in a total training dataset size of 293,109 patients (256,327
193 control patients; 36,782 HAI patients). Of these patients, 63% are from the Banner Health
194 dataset, 32% are from the eICU dataset, and 5% are from the MIMIC-III dataset. The majority
195 of these patients are treated under ICU or general ward settings. Between the two infection
196 cohort criteria (microbiology culture orders vs non-prophylactic antibiotics administration),

197 26,599 HAI patients are identified from microbiology lab and ICD-9 code, while 10,183
198 infection patients are identified from non-prophylactic antibiotic administration and ICD-9 code.

199 ***Model performance***

200 We compared machine learning algorithms in their ability to discriminate infection from control
201 patients using clinical data acquired up to one hour before clinical suspicion of infection. Our
202 results show that gradient boosting with two level decision trees yielded the best performance
203 with a mean AUC of 0.88, Specificity of 0.93 and Sensitivity of 0.54 at the break-even point
204 (where Sensitivity is approximately equal to positive predictive value (PPV), see Supplemental
205 Materials), Sensitivity of 0.80 and 0.64 respectively for when Specificity is 0.80 and 0.90 (Table
206 1: Xgboost). Abstained Adaptive Boosting with decision stump achieved a mean AUC of 0.85,
207 Specificity of 0.92 and Sensitivity of 0.47 at break-even point, Sensitivity of 0.73 and 0.54
208 respectively for when Specificity is 0.80 and 0.90 (Table 1: Abstained AdaBoost). Logistic
209 regression performs poorly compared with ensemble algorithms, with a mean AUC of 0.77,
210 Specificity of 0.91 and Sensitivity of 0.40 at break-even point, Sensitivity of 0.60 and 0.43
211 respectively for when Specificity is 0.80 and 0.90 (Table 1: Logistic Regression). These results
212 suggest that ensemble models are superior to linear models in predicting infection.

213 Next, we asked if ensemble models perform better than established empirical rules and clinical
214 scores in infection prediction. First, fever or high body temperature (>98.6 F) is one of the first
215 symptoms that lead to clinical suspicion of infection. Therefore, we compared temperature
216 measurements between the infection and control cohorts, and calculated the discriminative power
217 of temperature at one hour before infection suspicion. Temperature by itself has an AUC = 0.59

218 for detecting infection, which is far inferior to performance achieved with gradient boosting
219 (AUC = 0.88). Second, qSOFA – quick Sequential Organ Failure Assessment – was introduced
220 by the Third International Consensus Definitions for Sepsis and Septic Shock task force in 2016,
221 and is proposed as a quick assessment tool for identifying sepsis among patients with infection
222 [23]. Based on the Sepsis-3 criteria, we extracted Glasgow Coma Score, Systolic Blood Pressure,
223 and Respiratory Rate from the medical database, and derived qSOFA scores at one hour before
224 clinical suspicion of infection. In total 111,651 qSOFA scores were extracted, 22,460 from
225 infection cohort and 89,191 from control cohort (infection prevalence = 20.1%). We then
226 calculated the area under ROC curve of infection prediction by using qSOFA alone. qSOFA by
227 itself has an AUC = 0.59 when predicting infection at one hour before suspicion of infection. To
228 ensure a fair comparison with ensemble models, we re-trained the Gradient Boosting algorithm
229 using data from the subset of patient cohort that have qSOFA available. Gradient Boosting on the
230 patient subset achieves an AUC of 0.83 which is substantially better than the performance of
231 qSOFA. Overall our results suggest advantages of ensemble models over established clinical
232 methods in infection prediction.

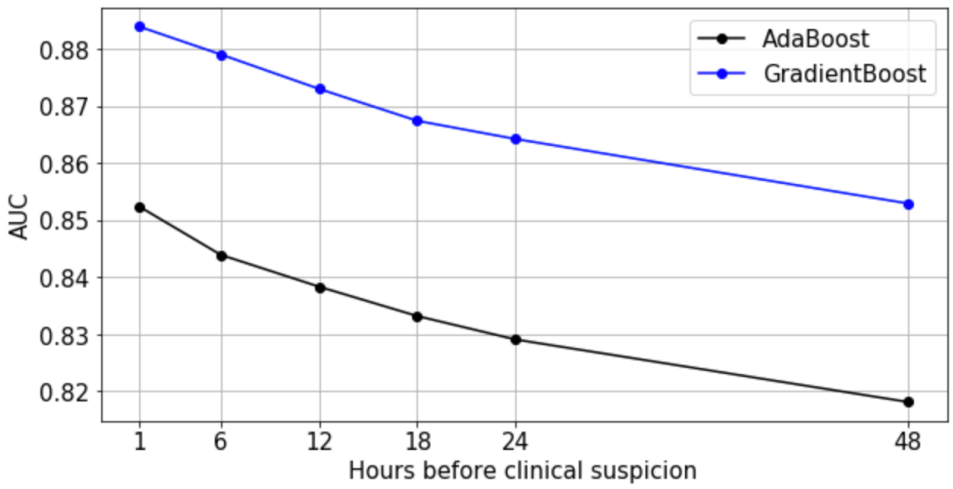
233 *Table 1: Performance of infection prediction at one hour before clinical suspicion of infection.*

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

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 one hour 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 one hour 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.

248 In addition, we investigated the infection prediction performance of ensemble models at earlier
249 time points. We applied the most interpretable model (Abstained AdaBoost) and the best
250 performing model (Gradient Boosting) to earlier observation windows to characterize predictive
251 performance over time using the full feature space (Figure 2). Despite degraded model
252 performance over time, Gradient Boosting maintains an AUC>0.85, while Adaptive Boosting
253 maintains an AUC>0.81 for 48 hours before clinical suspicion. These results support an assertion
254 that it is possible to predict hospital acquired infection earlier, up to 48 hours before clinical
255 suspicion of infection.



256 *Figure 2: Predictive performance of AdaBoost and GradientBoost models relative to time of clinical suspicion*

257 ***Model interpretation***

258 To better understand the biomarkers leveraged by the ensemble-based models, we first analyze
259 the AdaBoost algorithm with decision stumps since it is easier to interpret, and then contrast with
260 feature importance scores on the GradientBoost algorithm with decision trees using the SHAP
261 (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.

Table 2: Feature Importance Rankings from Abstained AdaBoost model (top 15). 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.*

Total Feature Importance		**Adjusted Feature Importance**	
Rank	Feature	Rank	Feature
1	Albumin	1	Albumin
2	Max(SpO2)	2	TIBC
3	pH	3	Fibrinogen
4	Min(SpO2)	4	Temperature
5	Temperature	5	ESR
6	Avg(SpO2)	6	PVRI
7	Var(SpO2)	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(SpO2)	12	CRP

13	FiO2	13	Ferritin
14	WBC	14	Neutrophils
15	Bicarbonate	15	Var(Temperature)

The learned risk functions behave in clinically interpretable ways. **Error! Reference source not found.** 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 3 plots) is unitless, but can be used to compare the relative importance of features (see 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

291 with inflammatory response: low TIBC (<240 mcg/dL; prevalence=3%), elevated Fibrinogen
292 (>325 mg/dL; prevalence=5%), and elevated ESR (> 45 mm/hr; prevalence=2%).

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

301 Although laboratory measurements play a significant role, the model also aggregates information
302 from a number of vital signs. The infection risk function based on temperature increases rapidly
303 above 37.8C, although this accounts for a small percentage of infection patients (5105 out of
304 40406 (~12.6%) of infection patients registered a fever >= 37.8C at the 1-hour window). For
305 controls, 5579 out of the 96505 control patients (~5.8%) exhibited a fever >= 37.8C. Infection
306 patients tend to have an elevated heart rate and macro variability, which is reported to be critical
307 for the diagnosis and prognosis of infection by many studies [28] [29]. For blood pressure,
308 patients tend to have a decreased blood pressure (systolic, diastolic, and mean), and this effect
309 was often selected by the classifier. Many trend variability features on vitals were selected
310 across temperature, heart rate, blood pressure, oxygen saturation(SpO2), and respiration, as the

infection cohort tends to exhibit a heavier right tail in feature variance measures. Changes in vital signs are also reported in the literature to accompany the development of infection [30] [31].

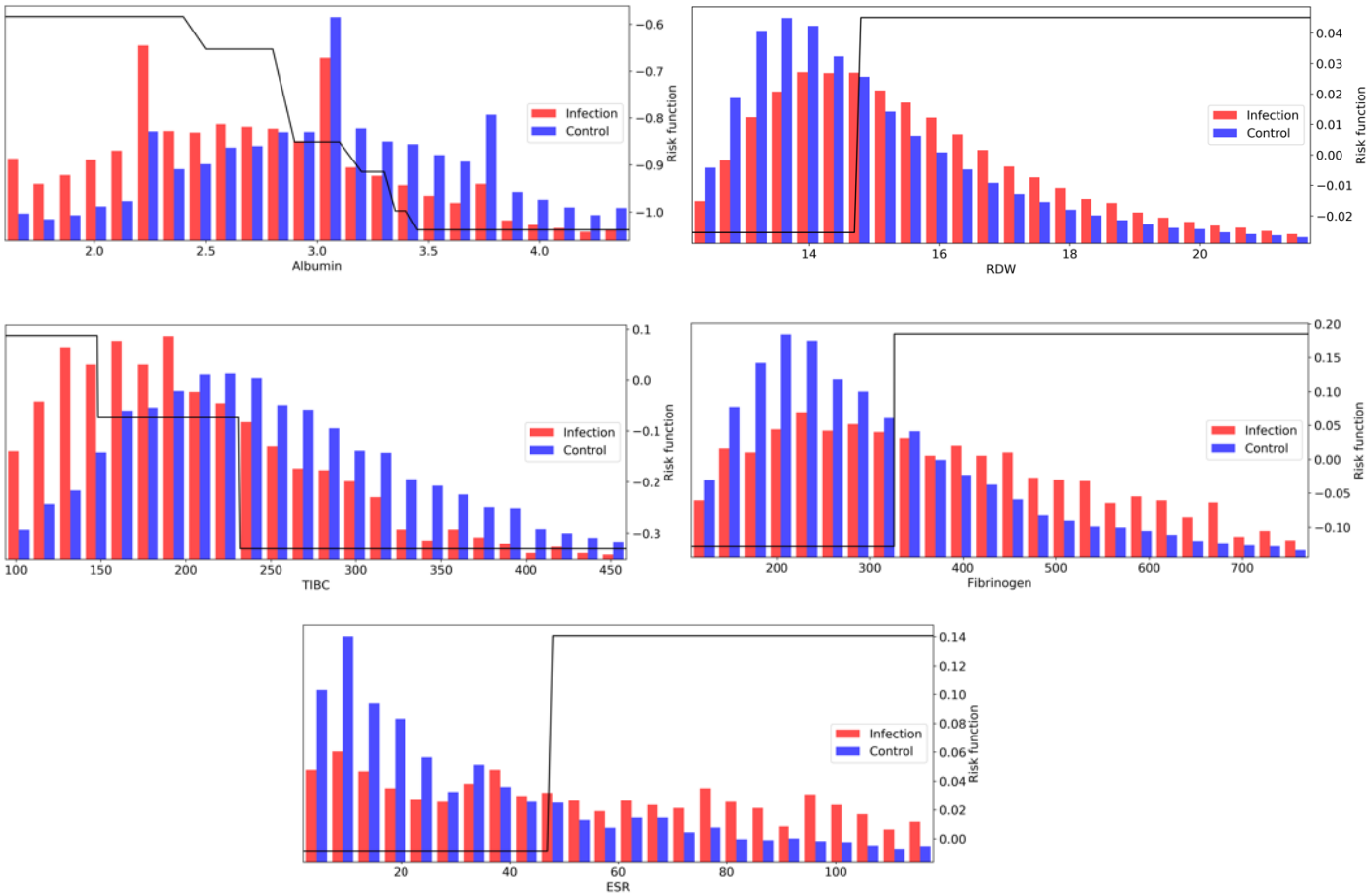


Figure 3: AdaBoost risk functions (black) for a subset of the most important laboratory measurements, along with population distribution underlays for infection (red) and control (blue) populations

We additionally applied SHAP analysis to extract feature importance rankings from the Gradient Boosting method (Figure 4). We have observed overlaps in the selected features between the more interpretable AdaBoost model and Gradient Boosting, such as Albumin, SpO2, Bicarbonate, Temperature, Lactate and BUN.

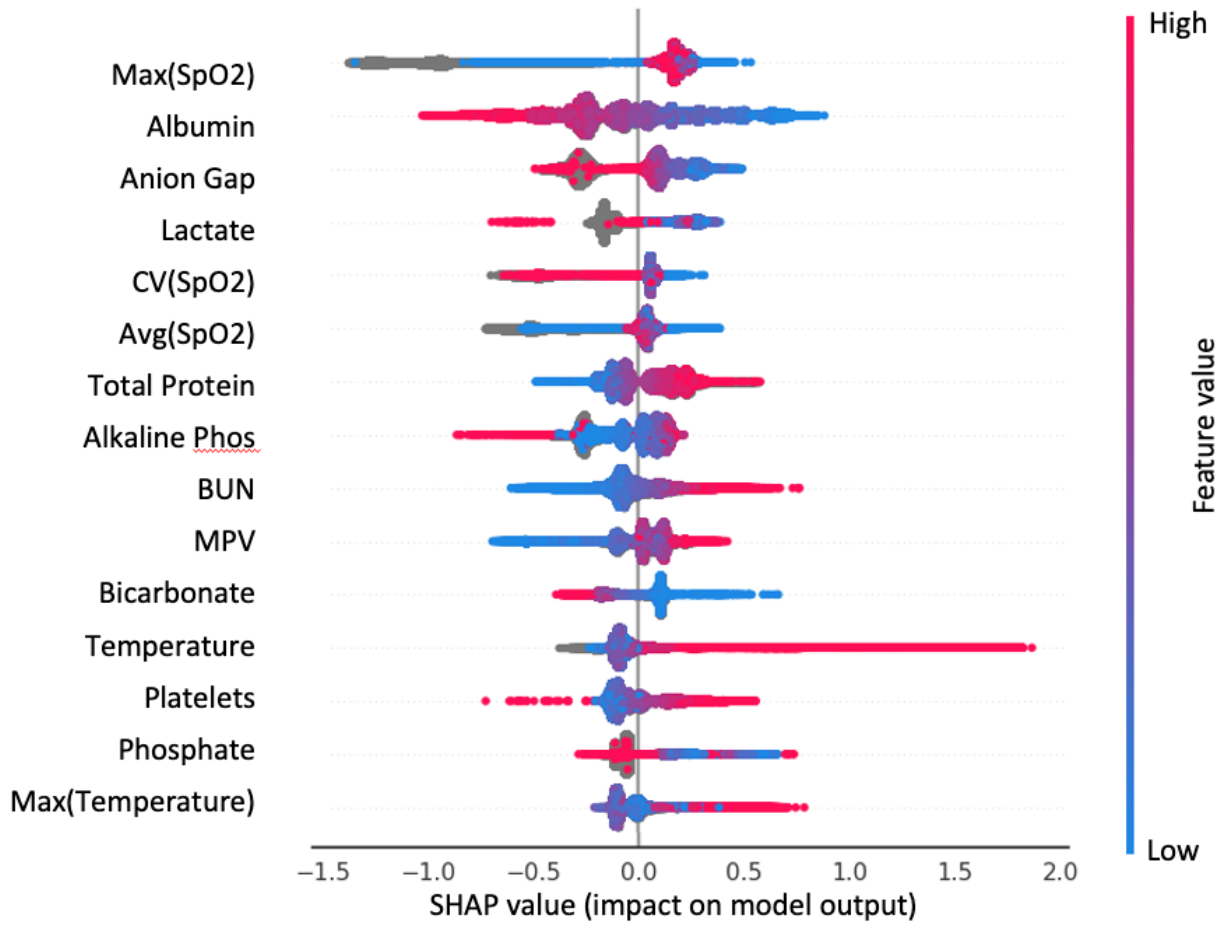


Figure 4: Top 15 important features of the GradientBoost model from SHAP analysis. Each dot is a patient; color indicates the value of the feature. SHAP value is on the x-axis: large positive value - feature contributes strongly to predict infection; large negative value – feature contributes strongly to predict control

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 Materials): 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 5). 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.

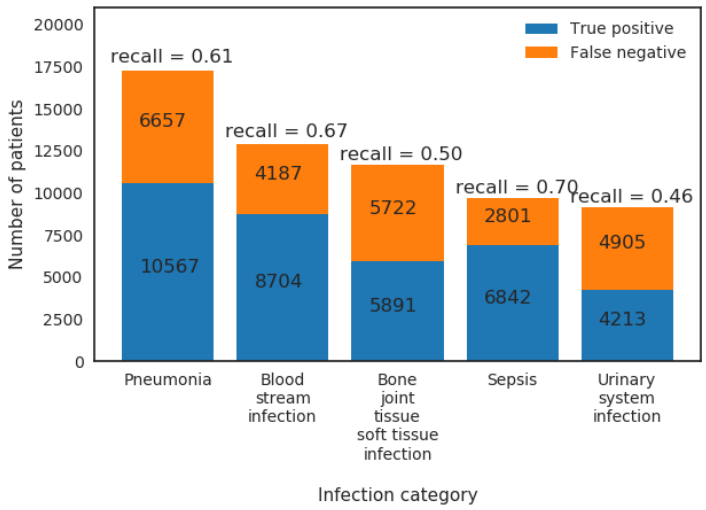


Figure 5: True positives and false negatives from GradientBoost model for the top five prevalence infection categories

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 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 6 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 7.

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.

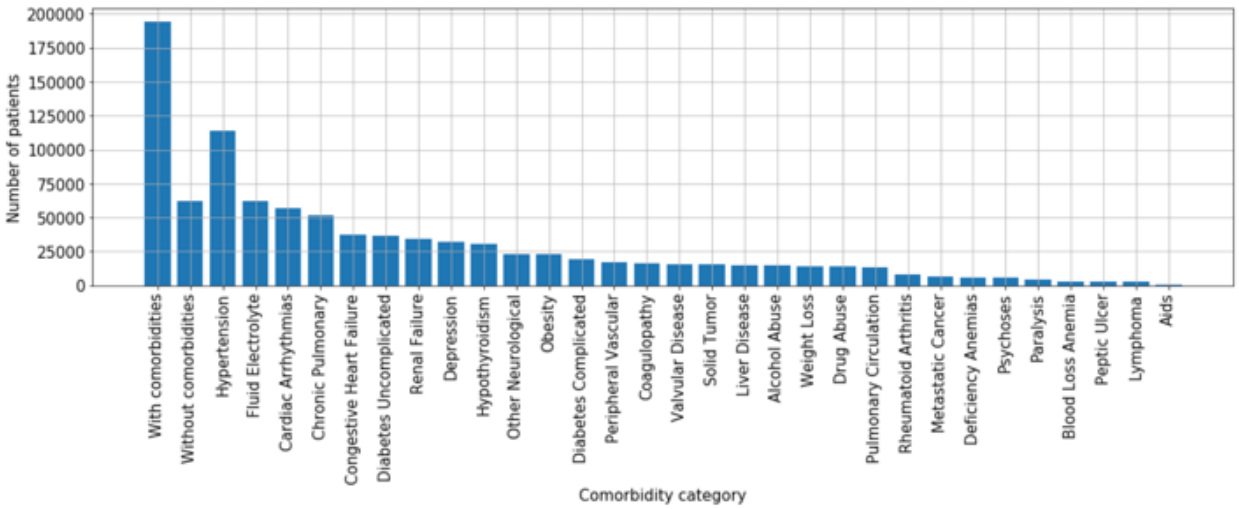


Figure 6: Comorbidity prevalence amongst control patients

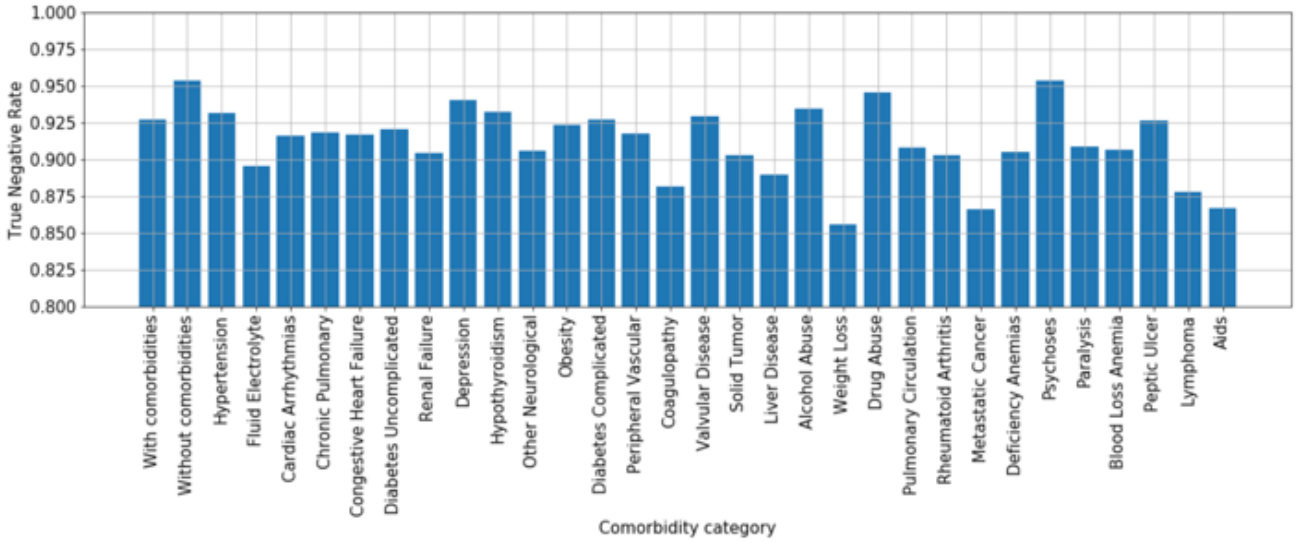


Figure 7: True negative rates (specificity) by comorbidity category. X-axis: "With comorbidities" - control patients with at least one comorbidity; "Without comorbidities" - control patients without any documented comorbidities; 30 comorbidity categories are ordered by prevalence shown in Figure 6 to highlight that the differences in True Negative Rate are not simple reflections of prevalence.

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-hours before clinical suspicion.

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] [29] [30] [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 “superfeature” exists that allows superior classification. This likely reflects at the same time the variable etiology of the HAI, which can be

386 of different natures (respiratory, blood stream infection, sepsis, etc.), the individual variability in
387 the response, and the multi-system nature of the effect of the infection on the patient's
388 physiology. On the other hand, it is still possible to obtain prediction performance that are
389 clinically viable with a reasonable number of clinical measurements. We have showed that with
390 a core set of 36 clinical measurements, the infection model performs at an AUC = 0.86 at one
391 hour before clinical suspicion of infection.

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

402 **CONCLUSION**

403 This study developed an algorithm for early identification of infection in hospitalized patients,
404 using machine learning applied to large retrospective hospital datasets. The model is able to
405 identify patients who are infected with reasonable performance up to 48 hours before clinical
406 suspicion of infection (AUC > .85). The trained models utilize ensembles of decision trees,

407 which are readily interpretable and provide ranked lists of feature importance. The primary
408 model leveraging all available (163) vital signs, laboratory measurements and demographics
409 achieves the best performance; however, a secondary model limited to the 36 most commonly
410 measured clinical measurements still achieves an AUC=0.86 at 1-hour before clinical suspicion.
411 The models compare favorably to established clinical rules and show high potential for real-
412 world hospital deployment as a clinical decision support aid.

413 **ACKNOWLEDGEMENTS**

414 This work is sponsored by the US Department of Defense (DoD), Defense Threat Reduction
415 Agency (DTRA) under project CB10560.

416 **LIST OF ABBREVIATIONS**

417 HAI: Healthcare-associated infection
418 CDS: Clinical decision support
419 qSOFA: quick Sequential Organ Failure Assessment
420 SHAP: Shapley Additive exPlanations
421 Spec: Specificity
422 TNR: True Negative Rate
423 AUC: Area under the ROC curve

424 **DECLARATIONS**

425 Ethics approval and consent to participate: The MIMIC-III project was approved by the
426 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.

Consent for publication: Not applicable

Availability of data and materials: 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.

Competing interests: Authors are employees of Philips Research, Banner Health and Massachusetts Institute of Technology.

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 design and interpretation of the study.

Authors' 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

449 identification, manuscript review and interpretation of results. All authors have read and

450 approved the manuscript.

REFERENCES

[1] C. f. D. C. a. Prevention, "2018 National and State Healthcare-Associated Infections Progress Report," 2019. [Online]. Available: http:///www.cdc.gov/hai/data/portal/progress-report.html.

[2] R. M. J. R. E. C. L. R. J. T. C. H. R. P. G. D. A. P. a. D. M. C. Klevens, "Estimating health care-associated infections and deaths in US hospitals, 2002," *Public health reports,* vol. 122, no. 2, pp. 160-166, 2007.

[3] M. A. K. E. S. S. S. H. K. K. E. J. S. N. V. J. B. e. a. Baker, "The impact of coronavirus disease 2019 (COVID-19) on healthcare-associated infections," *Clinical Infectious Diseases,* vol. 74, no. 10, pp. 1748-1754, 2022.

[4] R. D. M. M. T. A. E. P. D. J. L. T. a. W. B. MacArthur, "Adequacy of early empiric antibiotic treatment and survival in severe sepsis: experience from the MONARCS trial," *Clinical infectious diseases,* vol. 38, no. 2, pp. 284-288, 2004.

[5] R. A. A. D. S. E. P. M. M. L. A. A. X. L. P. a. J.-M. S. Ferrer, "Effectiveness of treatments for severe sepsis: a prospective, multicenter, observational study," *American journal of respiratory and critical care medicine,* vol. 180, no. 9, pp. 861-866, 2009.

[6] I. M. M. E. H. A. N. a. Y. Y. Longini Jr, "Containing pandemic influenza with antiviral agents," *American journal of epidemiology,* vol. 159, no. 7, pp. 623-633, 2004.

[7] P. J. C. A. G. a. R. M. W. Pronovost, "The wisdom and justice of not paying for “preventable complications”," *Jama,* vol. 299, no. 18, pp. 2197-2199, 2008.

[8] J. K. S. B. M. a. M. P. G. Long, "Antiviral agents for treating influenza," *Cleveland Clinic journal of medicine ,* vol. 67, no. 2, pp. 92-95, 2000.

[9] M. M. Churpek, A. Snyder, S. Sokol, N. N. Pettit and D. P. Edelson, "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.,* vol. 45, no. 11, p. 1805–1812, 2017.

[10] P. Bhattacharjee, D. P. Edelson and M. M. Churpek, "Identifying Patients With Sepsis on the Hospital Wards," *Chest,* vol. 151, no. 4, p. 898–907, 2017.

[11] C. A. Umscheid, J. Betesh, C. VanZandbergen, A. Hanish, G. Tait, M. E. Mikkelsen, B. French and B. D. Fuchs, "Development, Implementation, and Impact of an Autoated Early Warning and Response System for Sepsis," *J. Hosp. Med.,* vol. 10, no. 1, pp. 26-31, 2015.

[12] J. Holmes, G. Roberts, S. Meran, J. D. Williams, A. O. Phillips, W. Aki and S. Group, "Understanding Electronic AKI Alerts," *Kidney Int. Reports,* vol. 2, no. 3, p. 342–349, 2017.

[13] J. Holmes, T. Rainer, J. Geen, G. Roberts, K. May, N. Wilson and J. D. Williams, "Acute Kidney Injury in the Era of the AKI E-Alert," pp. 1-9, 2016.

[14] M. N. Jones, "The National Early Warning Score Development and Implementation Group," *Clin. Med. J. R. Coll. Physicians London,* vol. 12, no. 6, p. 501–503, 2012.

[15] A. McCoy and R. Das, "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.,* vol. 6, no. 2, p. e000158, 2017.

[16] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso and R. Das, "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.,* vol. 6, no. 2, p. e000158, 2017.

[17] S. Horng, D. A. Sontag, H. Y., Y. Jernite, N. I. Shapiro and L. A. Nathanson, "Creating an Automated Trigger for Sepsis Clinical Decision Support at Emergency Department Triage Using Machine Learning," *PLoS One,* vol. 12, no. 4, p. e0174708, 2017.

[18] A. P. T. S. L. L.-W. H. F. M. G. M. M. B. S. P. C. L. a. M. R. Johnson, "MIMIC-III, a freely accessible critical care database," *Scientific data,* vol. 3, no. 1, pp. 1-9, 2016.

[19] T. J. A. R. J. C. L. M. R. a. B. O. Pollard, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Scientific data,* vol. 5, p. 180178, 2018.

[20] B. Conroy, L. Eshelman, C. Potes and M. Xu-Wilson, "A dynamic ensemble approach to robust classification in the presence of missing data," *Machine Learning,* pp. 443-463, 2016.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[22] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, 2017.

[23] D. C. S. C. e. a. Singer M, " The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA,* vol. 315, no. 8, p. 801–810, 2016.

[24] M. W. G. P. J.P. Nicholson, "The role of albumin in critical illness," *British Journal of Anaesthesia,* vol. 85, no. 4, pp. 599-610, 2000.

[25] "High white blood cell count," *Mayo Clinic.*

[26] A. C. G. N. A. Pirillo, "HDL in infectious diseases and sepsis," *Handb Exp Pharmaccol,* vol. 224, pp. 483-508, 2015.

[27] M. H. F. K. a. W. Jelkmann, "Review: Role of Blood Platelets in Infection and Inflammation," *Journal of Interferon & Cytokine Research,* vol. 22, no. 9, p. 913–922, 2002.

[28] A. T. K. D. N. R. Z. a. A. J. S. S. Ahmad, "Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection," *Critical care (London, England),* vol. 13, no. 6, p. 232–232, 2009.

[29] A. S. S. M. M. a. G. L. A. S. N. Karmali, "Heart rate variability in critical care medicine: a systematic review," *Intensive care medicine experimental,* vol. 5, no. 1, p. 33–33, 2017.

[30] N. H. Z. Z. a. A. A. J. J. González Plaza, "Fever as an important resource for infectious diseases research," *Intractable & rare diseases research,* vol. 5, no. 2, p. 97–102, 2016.

[31] J. H. a. Y. Tokuda, "Changes in vital signs as predictors of bacterial infection in home care: a multi-center prospective cohort study," *Postgraduate Medicine,* vol. 129, no. 2, p. 283–287, 2017.

[32] H. Quuan and e. al., "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Medical care,* pp. 1130-1139, 2005.

[33] Q. W. X. B. L. C. H. L. Q. Z. Y. H. X. Y. L. Y. Z. a. A. L. Xu, ". "Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review," *Journal of Healthcare Engineering,* 2023.

ADDITIONAL FILES

Supplementary materials: Supplementary Material_v6.docx