American Journal of Infection Control 000 (2021) 1–8

Contents lists available at ScienceDirect

American Journal of Infection Control

journal homepage: www.ajicjournal.org

Major Article

A comparative analysis of machine learning approaches to predict *C. difficile* infection in hospitalized patients

Saarang Panchavati BS [1], Nicole S. Zelin MD [1], Anurag Garikipati MS, Emily Pellegrini MEng, Zohora Iqbal PhD *, Gina Barnes MPH, Jana Hoffman PhD, Jacob Calvert MSc, Qingqing Mao PhD, Ritankar Das MSc

Dascena, Inc., Houston, TX

Key Words:
Machine learning
Algorithm
Prediction
Clostridioides difficile
CDI
Electronic health record
XGBoost

A B S T R A C T

Background: Interventions to better prevent or manage *Clostridioides difficile* infection (CDI) may significantly reduce morbidity, mortality, and healthcare spending.

Methods: We present a retrospective study using electronic health record data from over 700 United States hospitals. A subset of hospitals was used to develop machine learning algorithms (MLAs); the remaining hospitals served as an external test set. Three MLAs were evaluated: gradient-boosted decision trees (XGBoost), Deep Long Short Term Memory neural network, and one-dimensional convolutional neural network. MLA performance was evaluated with area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, diagnostic odds ratios and likelihood ratios.

Results: The development dataset contained 13,664,840 inpatient encounters with 80,046 CDI encounters; the external dataset contained 1,149,088 inpatient encounters with 7,107 CDI encounters. The highest AUROCs were achieved for XGB, Deep Long Short Term Memory neural network, and one-dimensional convolutional neural network via abstaining from use of specialized training techniques, resampling in isolation, and resampling and output bias in combination, respectively. XGBoost achieved the highest AUROC.

Conclusions: MLAs can predict future CDI in hospitalized patients using just 6 hours of data. In clinical practice, a machine-learning based tool may support prophylactic measures, earlier diagnosis, and more timely implementation of infection control measures.

INTRODUCTION

Clostridioides difficile (*C. difficile*) infection (CDI) is the leading cause of hospital-acquired diarrhea and is associated with significant morbidity, mortality, and healthcare costs.[1–3] Over 450,000 cases of CDI and up to 29,000 CDI-related deaths are estimated to occur in the United States (US) every year.[4,5] Over the last decade, a decreasing trend in the incidence of CDI has been observed;[4] however, CDI remains a major clinical concern for hospitalized patients, especially among the growing demographic of geriatric patients.[2,6] In the US, excess annual healthcare spending related to CDI is estimated to be as much as $4.8 billion.[5] The average cost of treatment per case is approximately $4,000 with an average increase in hospital stay of 3.6 days.[7,8] Interventions to better prevent or improve the management of CDI may therefore be of compelling clinical and economic interest to clinicians and health systems.

Currently, there is no gold standard clinical risk assessment tool to predict in-hospital CDI. This study intends to fill this gap by exploring the feasibility of using machine learning algorithms (MLAs) to predict CDI across all hospitalized inpatients, providing early warning of a patient's risk of developing CDI. In clinical practice, such a tool may facilitate enhanced clinical monitoring, earlier diagnosis, and

* Address correspondence to: Zohora Iqbal, PhD, Dascena, Inc., 8515 Fannin St. Suite 110, Houston, TX 77054.
E-mail address: ziqbal@dascena.com (Z. Iqbal).

Funding/support: No external financial or material support was received to support this research.

Conflicts of interest: All authors who have affiliations listed with Dascena (Houston, Texas, U.S.A) are employees or contractors of Dascena.

Availability of data and materials**:** The data analyzed in this study was obtained from a proprietary longitudinal electronic health record (EHR) repository that includes over 700 hospitals located in the U.S. Requests to access the processed data and statistical information should be directed to Qingqing Mao, PhD, at qmao@dascena.com.

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ajic.2021.11.012

improved outcomes. From a public health perspective, such a tool would also empower clinical teams to implement appropriate infection control measures, like contact precautions, at earlier time points. Timely implementation of such infection control measures may help to minimize the spread of CDI to other vulnerable patients.[9]

In this study, we compare the performance of 3 different machine learning architectures in their ability to predict future in-hospital CDI after only 6 hours of a hospital stay: a gradient boosted decision tree ensemble implemented with XGBoost, a Deep Long Short Term Memory neural network (D-LSTM), and a 1-dimensional convolutional neural network (1D-CNN). We systematically evaluate the most effective training techniques to account for the relatively low prevalence of CDI in training data. We then demonstrate the generalizability of the best performing MLAs using a dataset distinct from the development dataset, drawn from diverse health systems.

MATERIALS AND METHODS

Datasets

Electronic health record (EHR) data for adult patients (≥ 18 years) were extracted from a proprietary national, longitudinal EHR repository derived from over 700 hospitals between 2007 and 2021. All data were collected passively and de-identified in compliance with the Health Insurance Portability and Accountability Act. As this project constituted "nonhuman subjects research" per *45 CFR 46,*[10] this project did not require institutional review board approval.

The comprehensive dataset was partitioned prior to any data analysis into a development dataset and a dataset to verify the generalizability of the final MLAs created. The latter dataset, which shall be hereafter referred to as the external dataset, was derived from geographically diverse health systems representing all 4 US census regions. Partitioning was performed at the level of the healthcare system, such that no systems included in the external dataset were included in the development dataset.

The development dataset was partitioned into an 80:20 split for training and initial testing of the MLAs. In other words, 80% of the development dataset was reserved for MLA training, and 20% was reserved as a hold-out test set to which the MLAs were not exposed during training. Initial performance results for the MLAs were evaluated on the hold-out test set. The performance of the best performing MLAs was then also evaluated on the external dataset. Characteristics of the CDI and non-CDI patients in the datasets were compared using 2-proportions z-tests.

Data Processing

For both the development and the external dataset, the following inclusion criteria were applied: (1) inpatient hospitalization with at least 6 hours of data recorded; (2) age ≥ 18 at time of inpatient hospitalization; and (3) recorded measurement for at least one vital sign in the first 6 hours of hospitalization. We excluded patient visits with a positive laboratory test for CDI prior to 6 hours (the prediction time for the MLAs) to ensure that predictions were only generated in patients not yet diagnosed with CDI. We extracted demographics, vital signs, laboratory tests, other clinical measurements, medication treatments, and medical history to provide as inputs to the MLAs (Table 1). Inputs were selected based on *a priori* relevance to CDI prediction or previously reported associations with CDI risk in the literature, providing a guide for commonly evaluated measures when determining the presence of CDI.[11,12] During an initial data evaluation, a broad set of potentially relevant features was extracted. A subset of features with the highest prevalence in the dataset was then selected to maximize input availability.

Table 1
Data inputs provided to all machine learning algorithms

Demographics	
Age	Sex
Race	
Vital Signs	
Systolic blood pressure	Respiratory rate
Diastolic blood pressure	Peripheral oxygen saturation
Heart rate	Temperature
Other Clinical Measurements	
Body mass index (BMI)	
Laboratory Tests	
Sodium	Hemoglobin
Creatinine	White blood cell count
Blood urea nitrogen (BUN)	Platelet count
Glucose	Glycated hemoglobin (A1c)
Aspartate aminotransferase	Total bilirubin
Alanine aminotransferase	Direct bilirubin
Homocysteine	
Active Medication Treatment	
Proton pump inhibitors	Antibiotics (any)
Histamine H2 antagonists	Nonsteroidal anti-inflammatory drugs (NSAIDs)
Comorbid Medical Conditions	
Hepatic cirrhosis	Current tobacco use
Inflammatory bowel disease	

Patients were considered to be currently receiving a medication if the medication was ordered within the 30 days prior to hospitalization or within the first 6 hours of hospitalization. Comorbid medical conditions may have been recorded in the EHR at any time point prior to prediction generation.

All MLAs incorporated binary features, such as active medication treatment or presence of medical comorbidities, in the same fashion: a patient encounter was considered to either have the feature present prior to prediction or absent prior to prediction. For continuously measured features, such as vital sign and laboratory measurements, XGBoost used the last measured value, as well as summary statistics (mean, standard deviation) of all values measured during the first 6 hours of hospital care. For continuously measured features, the neural networks used the raw time series data. For the XGBoost model, a null value was reported if a feature was not measured for a patient, and the null value was later implicitly handled by the MLA. For the neural network models, a value of -1 was imputed for missing values to indicate missingness.

Predictions were made at 6 hours into the hospital stay because 84% of patients who developed CDI in the development dataset received diagnostic confirmation via laboratory testing after 6 hours. This time point also enabled the collection of sufficient clinical data for MLAs, like the neural networks, which require substantial training data.

Gold standard

Patient encounters were considered positive for CDI if a positive laboratory test result for CDI was detected via analysis of structured laboratory data, or through natural language processing of clinical notes. Patients with any positive molecular test for CDI, including polymerase chain reaction and enzyme immunoassay tests, were included in the positive class, consistent with current Infectious Disease Society of America guidelines for the diagnosis of CDI.[13] A positive test result could be returned at any point after 6 hours during the hospitalization to satisfy the criterion for a CDI positive encounter. All patient encounters not meeting this criterion were considered negative, including patient encounters with negative CDI test results or with no CDI testing performed.

Machine learning models

Three different MLAs were developed to predict CDI using data available 6 hours into the inpatient stay. For the development of the

models, we labeled patient encounters as uniquely positive or negative per the gold standard. CDI encounters as defined by the gold standard were included as the positive class. Non-CDI patient encounters were included as the negative class. The 3 types of models selected were gradient-boosted decision tree ensemble implemented via XGBoost, a 1D-CNN, and a D-LSTM. These models were chosen to represent a variety of classical machine learning and deep learning methods. XGBoost, a classical method, has achieved high performance across a range of health-related prediction tasks.[14–16] As deep learning approaches, convolutional neural networks (CNN) and long short-term memory networks can effectively incorporate time series data to predict health outcomes.[17,18]

The gradient-boosted decision tree classifier was implemented by the XGBoost method in Python.[19] Gradient boosting combines results from various decision trees to generate prediction scores. Each decision tree splits the patient population into successively smaller groups, and each branch of the tree divides patients into 2 groups according to their covariate value. Each patient encounter was ultimately represented in one "leaf" of the tree. Each encounter in the same leaf was predicted to have the same risk of CDI. As part of model training and development, 5-fold cross-validation was performed to select the optimized hyperparameters for the XGBoost model. XGBoost was tuned to the following hyperparameters, which were used when evaluating the performance of the best performing XGBoost model (described below) on the development hold-out test set and the external dataset: regularization term of 10, 250 estimators, a maximum tree depth of 10, and a learning rate of 0.1.

The D-LSTM model uses a stacked Bidirectional LSTM followed by an LSTM layer that extracts both short and long term trends across features and timesteps. An LSTM allows the model to identify relationships between different timesteps across the same input, providing the ability to have features persist through the model and identify lagging trends in the data. The outputs from the LSTM layers are fed into successively dense layers where further abstractions from the time series data are used to develop an understanding of how features through time can map to a patient having CDI or not. The 1D-CNN model is a variant of a CNN which applies a convolution operation to each feature input over the 6-hour time period for which data was used. The convolution identifies local trends in the data over time for each of the individual features. Following the convolutions, a max pooling layer and a dense layer are applied to identify global features across all of the inputs, allowing the model to gain an understanding on how different inputs are related. The final layer of the 1D-CNN then maps each patient's data to a value in order to classify a patient as having or not having CDI. Both neural network architectures concatenate constant-valued features, such as age and sex, with the outputs from the LSTM or convolution layers to further augment the predicted outcome. Both neural networks were trained with an early stopping parameter to minimize the possibility of overfitting and reduce computation time. The architectures of the neural network models are presented in Supplementary Figure 1.

The incidence of CDI across the general inpatient population has been reported to be under 2%.[20] Low prevalence outcomes present a training challenge for machine learning classifiers, as the minority class has far fewer examples in the training data from which the model can learn the characteristics of the class. Multiple techniques have been developed to enhance the training of MLAs for the prediction of rare outcomes. To optimize the performance of each of the MLAs selected for this prediction task, we systematically evaluated the impact of applying the following training techniques alone and in combination: positive weight scaling (as part of hyperparameter tuning), resampling the positive class so that CDI achieves 20% prevalence, and output bias initialization.[21–23] The output bias method was only evaluated for the neural networks, as the bias term adjusts the initial output of a network to better reflect the imbalanced distribution of the data. This adjustment enables the network to avoid spending training time learning that positive CDI samples are highly unlikely in the data. As XGBoost builds a tree-based model, no bias shifting is necessary to fit the model to the training data. Resampling was only performed routinely on the training dataset to ensure the models had sufficient minority class data to learn during the training process. Thus, even when this training technique was evaluated, no resampling was performed on the hold-out test or external datasets. The prevalence of the hold-out test and external datasets thus remained as described in Supplementary Figure 2, to best approximate model performance on real-world data. When resampling was not one of the training techniques under investigation, no resampling during model training occurred. Performance of each MLA with the specified training technique(s) was evaluated using the area under the receiver operating characteristic curve (AUROC), and the combination of techniques resulting in the highest AUROC for each MLA was selected as the best performing version of the MLA.

The most important input features for each of the best performing MLAs were then evaluated on the development hold-out test set. For the XGBoost model, a Shapley Additive Explanations (SHAP) analysis was performed to evaluate the inputs that the MLA determined to be the most important features in making predictions.[24] A SHAP summary plot was generated, which ranks the features in terms of descending importance and visually displays relationships between the directionality of features (e.g., high or low; present or absent) and predictions of risk. A mean feature importance plot was also generated to summarize the average importance of the most important features, regardless of directionality. The mean feature importance was determined by taking the magnitude of each feature's SHAP values for the entire dataset. For the neural networks, a similar analysis was performed to generate a mean feature importance plot using the magnitude of SHAP values across the entire time series of each of the features. Due to the complexity of the neural networks and the computational difficulty of calculating the SHAP values on the full development hold-out test set, the SHAP values were calculated based on a subset of the data. For the background expectation, 1,000 training samples were used, and 200 testing samples were used to generate the SHAP summary plot to explain their outputs. Given the greater analytic complexity of neural networks, a SHAP summary plot showing the directionality of relationships between feature values and estimations of risk cannot be generated. For both the XGBoost and neural network models, feature importance plots were limited to the top ten most important features used to generate predictions.

The performance of the best performing XGBoost, D-LSTM and 1D-CNN models were first evaluated on the development hold-out test set in terms of AUROC, sensitivity, specificity, positive and negative likelihood ratios (LR+, LR-), and the diagnostic odds ratios (DOR). For each MLA, the clinical operating points at which sensitivity, specificity, DOR, LR+ and LR- were evaluated was selected as the point at which both sensitivity and specificity were maximized. This clinical operating point was determined by identifying the operating point that gave the maximum geometric mean of sensitivity and specificity. To further validate the MLAs and determine whether their performance can generalize to health systems from which EHR data was not used in development, performance of the best performing models was evaluated on the external dataset using the same metrics.

RESULTS

After application of the study inclusion criteria, the development dataset contained 13,664,840 total inpatient encounters and 80,046 encounters with a positive CDI test (Supplementary Fig 2). In the 20% hold-out test set, 2,732,968 inpatient encounters and 16,009 CDI encounters were included. The prevalence of CDI was therefore

Table 2
Demographic information for the study sample in the development training data, the development hold-out test set and the external dataset

		Development Hold-Out Test Set			External Dataset		
		Patients with CDI	Patients without CDI	*P*-value	Patients with CDI	Patients without CDI	*P*-value
Age	<30	611 (3.82%)	326489 (12.02%)	< .001	1013 (14.25%)	288847 (25.29%)	< .001
	30-49	2094 (13.08%)	3607013 (22.34%)	< .001	1091 (15.35%)	167061 (14.63%)	.086
	50-59	2476 (15.47%)	429695 (15.82%)	.228	1414 (19.9%)	197947 (17.33%)	< .001
	60-69	3434 (21.45%)	487596 (17.95%)	< .001	1564 (22.01%)	192213 (16.83%)	< .001
	70-80	3759 (23.48%)	480870 (17.70%)	< .001	1663 (23.40%)	157499 (13.79%)	< .001
	80+	3635 (22.71%)	385296 (14.18%)	< .001	362 (5.09%)	138414 (12.12%)	< .001
Sex	Female	9211 (57.54%)	1578850 (58.11%)	.142	3973 (55.9%)	686342 (60.1%)	< .001
	Male	6791 (42.42%)	1136695 (41.84%)	.136	3130 (44.04%)	455159 (39.86%)	< .001
	Unknown	7 (0.04%)	1414 (0.05%)	.645	4 (0.06%)	480 (0.04%)	.559
Race	White	13283 (82.97%)	2140763 (78.79%)	< .001	1057 (14.87%)	163400 (14.31%)	.176
	Black	1911 (11.94%)	359207 (13.22%)	< .001	324 (4.56%)	65082 (5.70%)	< .001
	Asian	103 (0.64%)	31081 (1.14%)	< .001	5081 (71.49%)	792885 (69.43%)	< .001
	Other/Unknown	712 (4.45%)	185908 (6.84%)	< .001	645 (9.08%)	120614 (10.56%)	< .001
Ethnicity	Hispanic	630 (3.94%)	164492 (6.05%)	< .001	557 (7.84%)	93409 (8.18%)	.294
Comorbid	HIV/AIDS	943 (5.89%)	6027 (0.22%)	< .001	74 (1.04%)	2313 (0.20%)	< .001
Diseases	Peripheral vascular disease	50 (0.31%)	6985 (0.26%)	.169	239 (3.36%)	12849 (1.13%)	< .001
	Chronic heart failure	769 (4.80%)	4953 (0.18%)	< .001	41 (0.58%)	1511 (0.13%)	< .001
	Chronic kidney disease	140 (0.87%)	20330 (0.75%)	.0647	152 (2.14%)	6433 (0.56%)	< .001
	Hepatic cirrhosis	702 (4.39%)	25901 (0.95%)	< .001	227 (3.19%)	8663 (0.76%)	< .001
	History of organ transplant	978 (6.11%)	27888 (1.03%)	< .001	1163 (16.36%)	9310 (0.82%)	< .001
	Diabetes mellitus	838 (5.23%)	38670 (1.42%)	< .001	9 (0.13%)	2226 (0.19%)	< .001
	COPD	268 (1.67%)	26351 (0.97%)	< .001	71 (1.0%)	4071 (0.36%)	< .001
	Active cancer	808 (5.05%)	39706 (1.46%)	< .001	23 (0.32%)	15028 (1.32%)	< .001
	HTN	498 (3.11%)	57714 (2.12%)	< .001	90 (1.27%)	8519 (0.75%)	< .001
	Dementia	287 (1.79%)	22481 (0.83%)	< .001	149 (2.1%)	7701 (0.67%)	.193
	IBD	176 (1.10%)	5295 (0.19%)	< .001	257 (3.62%)	544 (0.05%)	< .001
	CDI within the past year	1047 (6.54%)	40164 (1.48%)	< .001	376 (5.29%)	7781 (0.68%)	< .001

P-values were generated using a 2 proportion z-test. Abbreviations: Acquired immunodeficiency syndrome (AIDS); Chronic obstructive pulmonary disease (COPD); Hypertension (HTN); Human immunodeficiency virus (HIV); Inflammatory bowel disease (IBD); C. difficile infection (CDI).

0.586% in both the overall development dataset and the hold-out test. The external dataset consisted of 1,149,088 inpatient encounters, including 7,107 encounters with a positive CDI test (Supplementary Fig 2). The prevalence of CDI in the external dataset was 0.618%.

Demographics were evaluated on the development hold-out test set and on the external dataset (Table 2). Patients who developed CDI were, on average, likely to be older. In the development hold-out test set, median age among patients without CDI was 59 years (interquartile range (IQR): 40, 74), and median among those with CDI was 68 (IQR: 55, 79). In the external dataset, median age among patients without CDI was 58 years (interquartile range (IQR): 37, 73), and median among those with CDI was 67 (IQR: 54, 79). Patients who developed CDI were also more likely to have a medical history of heart failure, diabetes mellitus, inflammatory bowel disease, active cancer, organ transplant, human immunodeficiency virus (HIV) infection or acquired immunodeficiency syndrome (AIDS), and prior CDI within the last year (Table 2).

For the XGBoost model, abstaining from the use of any specialized training techniques to accommodate for the low prevalence of CDI yielded the highest performance in terms of AUROC (Table 3). For the 2 neural network models, slight differences were seen when comparing the techniques that enabled the best performance: resampling in isolation for the D-LSTM model, and resampling and output bias in combination for the 1D-CNN model. These versions of each MLA were then selected as the best performing MLAs for all subsequent experiments.

The SHAP analysis for the best performing XGBoost model revealed that age was the most important feature in generating predictions, with clinical measurements (eg, sodium, BMI, white blood cell count, bilirubin, heart rate, diastolic blood pressure) and active medication treatment with antibiotics or proton pump inhibitors (PPIs) also among the most important features (Supplementary Fig 3). Age was the most important feature for the D-LSTM model, and the second most important feature for the 1D-CNN model. Active antibiotic and PPI treatment remained among the most important features for the best performing D-LSTM and 1D-CNN. Glycated hemoglobin (A1c) and White race were identified as important features for each neural network, but not for the XGBoost model.

The best performing MLAs were then evaluated on the development hold-out test set, and the external dataset. Receiver operating characteristic (ROC) curves for the MLAs are presented in Figure 1. The ROC curves revealed that the XGBoost model achieved marginally higher sensitivities across a range of specificities compared to the neural networks. The ROC curves of the D-LSTM and 1-D LSTM were very similar, and the clinical operating points of the neural network models were also similar.

Table 3
Area under the receiver operating characteristic curve (AUROC) achieved by XGBoost gradient-boosted decision tree, Deep Long Short Term Memory neural network (D-LSTM) and one-dimensional convolutional neural network (1D-CNN) classifiers in combination with techniques to adjust for low prevalence of predicted outcome

Technique(s) to adjust for low prevalence of predicted outcome during training	XGBoost	D-LSTM	1D-CNN
No Positive Scaling or Other Training Technique	**0.815**	0.612	0.796
Positive Scaling	0.808	0.799	0.807
Resampling	0.792	**0.804**	0.800
Output Bias	-	0.782	0.796
Positive Scaling + Resampling	0.807	0.803	0.799
Positive Scaling + Output Bias	-	0.797	0.809
Resampling + Output Bias	-	0.801	**0.810**
Positive Scaling + Resampling + Output Bias	-	0.803	0.809

Results were generated on the development hold-out test set. The highest AUROC achieved for each machine learning architecture is bolded.

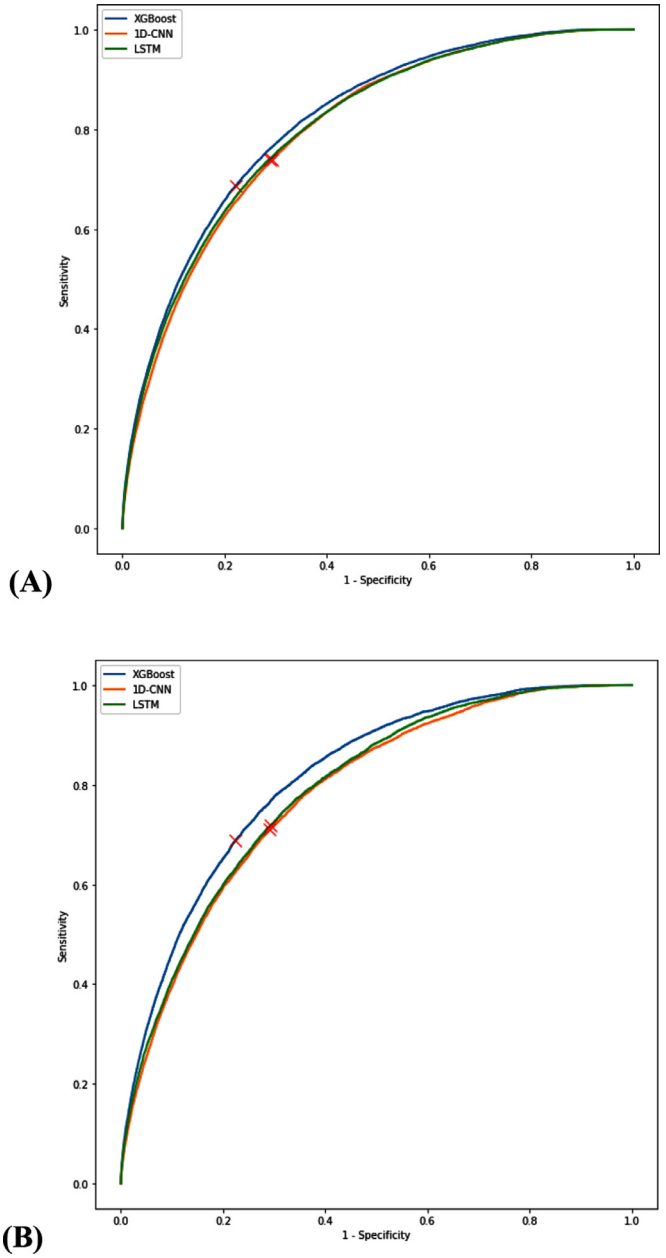

Fig 1. Receiver operating characteristic (ROC) curves of the best performing XGBoost gradient-boosted decision tree, Deep Long Short Term Memory neural network (D-LSTM) and one-dimensional convolutional neural network (1D-CNN) models for predicting *C. difficile* infection using the first 6 hours of inpatient hospitalization data. ROC curves were separately derived on (A) the development hold-out test set and (B) the external dataset. The clinical operating point for each model was selected via the maximum geometric mean of sensitivity and specificity, and was designated with a red x.

The best performing MLAs were also evaluated in terms of sensitivity, specificity, DOR, LR+ and LR- (Table 4). Overall, the XGBoost model achieved the highest performance in terms of AUROC. When operating points were selected via geometric mean to balance sensitivity and specificity, the neural networks demonstrated statistically significantly higher sensitivity than the XGBoost model. However, the operating point selected for the XGBoost model resulted in statistically significantly greater specificity than the D-LSTM and 1D-CNN.

The XGBoost MLA maintained its predictive performance in the external dataset. In comparison, the neural network models demonstrated decreased performance on the external dataset. In addition, for several performance metrics where XGBoost demonstrated non-statistically significant improvements over the neural network models on the hold-out test set, these improvements became statistically significant on the external dataset.

DISCUSSION

Risk stratification of individual patients early in a hospital admission may assist in the prevention of CDI in high risk patients.[25,26] Using real-world EHR data, we developed 3 different MLAs to predict CDI at any point during an inpatient stay based on only 6 hours of data. In clinical practice, use of a machine learning-based CDI prediction tool may enable patients to benefit from increased monitoring and treatment earlier in their disease course and facilitate timely implementation of appropriate infection control practices.

Our results demonstrate that MLAs can predict CDI with excellent discrimination (AUROC > 0.8).[27] Many of the important features used by the models to predict CDI were similar across MLAs, and have previously been identified as risk factors for CDI.[11,12] The highest performing MLA in terms of AUROC values was the XGBoost model, while the neural networks achieved higher sensitivities at the optimized operating points. The ROC curve demonstrates the range of possible sensitivity and specificity values at which an MLA (or other screening test) may operate.[27] A different operating point along the XGBoost curve could therefore have intentionally been selected so that the XGBoost model outperformed the D-LSTM and 1D-CNN in terms of sensitivity. Indeed, at any fixed specificity, the XGBoost model achieved a higher sensitivity than either neural network, as shown by the ROC curves. However, the fact that our standardized approach to selection of an operating point which balanced out sensitivity and specificity yielded an operating point with a lower sensitivity but higher specificity than the neural networks indicates that the XGBoost model learned to predict a more skewed distribution, with a greater imbalance between the majority negative class and minority positive class, than the neural networks. This observation is interesting in light of the superior predictive performance of XGBoost when there is no use of a specialized training technique to adjust for class imbalance compared to XGBoost's performance when such training techniques are used. The XGBoost model may therefore have learned from the class imbalance in training, such that the operating point reflected this MLA having more effectively "learned" to correctly identify encounters without the disease, the very definition of specificity.[28]

In this study, XGBoost achieved comparable predictive performance for CDI to the performance of more complex architectures, like D-LSTM and 1D-CNN. A variety of neural network architectures have been used to model time series data, in part because they can incorporate long input sequences and learn from time series with missing data.[29] In particular, recurrent neural network architectures have been applied to many clinical time series classification tasks.[30-32] However, in this instance, neither of the neural network models outperformed XGBoost. This may be because predictions were made on the basis of relatively short (6 hour) windows of time series data; neural networks may exhibit an advantage over XGBoost for longer time series. While XGBoost achieved comparable performance for the specified task on the available datasets, more complex neural networks may still provide a predictive advantage for different tasks or on different populations.[33-35]

The XGBoost MLA maintained its performance when examined on both the development hold-out test set and the external dataset. The performance of both neural networks decreased slightly when examined on the external dataset versus on the development hold-out test set. Increased MLA complexity is associated with greater model variance — and thus greater potential for overfitting to training data, which can impede generalizability to different datasets.[36] Both 1D-CNN and LSTM are far more complex and thus, more prone to

Table 4
Predictive performance of the best performing XGBoost gradient-boosted decision tree, Deep Long Short Term Memory neural network (D-LSTM) and 1-dimensional convolutional neural network (1D-CNN) models on the hold-out test set and external dataset

	Development Hold-Out Test Set			External Dataset		
	XGBoost	D-LSTM	1D-CNN	XGBoost	D-LSTM	1D-CNN
AUROC (95% CI)	**0.815** (0.812 - 0.819)	0.804 (0.801 - 0.807)	0.810 (0.807 - 0.813)	**0.815** (0.810 - 0.819)	0.789 (0.784 - 0.793)	0.781 (0.776 - 0.785)
Sensitivity (95% CI)	0.686 (0.679 - 0.694)	**0.739** (0.731 - 0.746)	**0.739** (0.733 - 0.746)	0.687 (0.678 - 0.698)	**0.718** (0.707 - 0.728)	0.709 (0.697 - 0.717)
Specificity (95% CI)	**0.779** (0.778 - 0.779)	0.713 (0.713 - 0.714)	0.724 (0.723 - 0.724)	**0.776** (0.775 - 0.777)	0.708 (0.707 - 0.709)	0.706 (0.705 - 0.706)
DOR (95% CI)	**7.710** (7.440 - 7.978)	7.043 (6.767 - 7.328)	7.427 (7.183 - 7.683)	**7.628** (7.291 - 8.032)	6.169 (5.852 - 6.472)	5.835 (5.507 - 6.097)
LR+ (95% CI)	**3.103** (3.067 - 3.135)	2.578 (2.552 - 2.605)	2.677 (2.487 - 2.533)	**3.071** (3.026 - 3.121)	2.458 (2.419 - 2.491)	2.408 (2.367 - 2.439)
LR- (95% CI)	0.402 (0.393 - 0.412)	0.366 (0.356 - 0.377)	**0.360** (0.352 - 0.369)	0.403 (0.389 - 0.415)	**0.398** (0.385 - 0.413)	0.412 (0.400 - 0.430)

The highest performance metric achieved in each dataset is bolded. Abbreviations: Area under the receiver operating characteristic curve (AUROC); Diagnostic odds ratio (DOR); Likelihood ratio (LR).

overfitting than XGBoost. In addition, the XGBoost model architecture has a maximum depth which limits the impact of each feature on each tree the model builds.[19] This functionality can lead to overall better generalizability to hold-out test and external datasets. Model complexity may thus impact the transferability of MLAs, such that more computationally intense MLAs may benefit from dataset-specific customization encompassing the entire development process, including feature selection, hyperparameter tuning and training.

This study further explored the impact on MLA performance of varied training techniques to account for the low prevalence of CDI. For both neural network models, resampling the positive class to artificially boost CDI prevalence in the training data significantly improved MLA performance. When resampling occurred, the neural networks were trained using batch gradient descent, and thus were able to "see" significantly more of the positive class through the training process than without resampling, likely enhancing the efficacy of the training process. The best performing D-LSTM MLA only used resampling, while the best performing 1D-CNN MLA used both resampling and an output bias initialization. This difference may be due to the fact that the modified output bias initialization helped nudge the 1D-CNN out of a local minimum, while there no such impact would occur for the D-LSTM. The best performing XGBoost model required no adjustment techniques to account for the class imbalance, including hyperparameter tuning which led to the positive weight scaling being unitary. Resampling did not improve XGBoost's performance, unlike with the neural networks. The lack of improvement may be attributable to the nature of the MLA. Tree-based models rely on splits in data based on strict thresholds. For a positive class with a range of feature values,is tree-based models may benefit less from upsampling since the positive class still spans a wide range of values, therefore preventing the model from learning quality splits for the trees.

To reduce the risk of infection-related adverse patient outcomes and healthcare costs, a number of conventionally hand-tabulated scoring systems have been developed to evaluate various outcomes associated with CDI.[37-39] These systems include the CARDS score for CDI-associated hospital mortality,[37] the ATLAS score for patient responsiveness to CDI therapy,[38] and Horn's index for CDI patient prognosis.[39] These existing scores have several limitations, in addition to their development for purposes other than predicting the future development of CDI. Some scores were developed using small, administrative databases[39] or using data that did not account for individual-level prognostic markers,[37] meaning that resulting scores may not be widely generalizable. There is a continued need for reliable, and accurate risk stratification tools to identify patients with an increased likelihood of developing CDI.[40]

Risk stratification models have also been developed to explicitly predict which patients are at risk for developing CDI, including models that use EHR data[40-44] and/or machine learning methods.[40-42] Some tools have been limited in the scope - e.g., only predicting recurrent CDI[41] - while other tools have been developed in specialized patient populations, like post-colectomy patients.[42] A smaller number of these models have utilized EHR data available early in a hospital admission,[25,44] with some still requiring manual tabulation.[25,44] However, none of the experimental risk prediction tools for CDI have yet been adopted in widespread clinical practice.

Compared to previous literature describing CDI prediction tools, our study has several strengths. The development dataset included a large, robust sample of general inpatients from across the US for training, and a separate hold-out test set for initial validation. In addition, this study incorporated a separated external dataset composed of distinct, geographically diverse US health systems. The maintenance of performance shown by the XGBoost model, and the largely consistent performance of the neural networks (with only minor decreases in predictive performance) supports the MLA's generalizability. All MLAs in this study were designed to automatically incorporate EHR data to generate risk scores, decreasing barriers to adoption (such as the requirement for time or resource investment by clinicians) and enhancing the potential for seamless integration into the clinical workflow. The best performing MLAs described in our current work offer strong predictive performance with the capacity to easily tailor operating points to the clinical needs or preferences of specific healthcare systems or clinical practice contexts. From a machine learning perspective, this study also included a rigorous investigation of the most appropriate and highest performing MLA architectures and training techniques. Previous studies exploring MLAs for applications related to CDI have not shown consistently strong performance across multiple architectures. Marra et al. evaluated ten architectures for the prediction of CDI in symptomatic hospitalized patients, with logistic regression, random forest and naïve Bayes models demonstrating the best performance.[45] However, no MLA achieved an AUROC greater than 0.60. Multilayer Perceptron and Radial Basis Function (RBF) neural networks achieved AUROC of only 0.575 and 0.583, respectively. Li et al. reported limited discrimination (AUROC of 0.69) with a logistic regression model to predict CDI severity.[46] Oh et al. achieved higher AUROCs (0.75, 0.82) with L2 regularized logistic regression models predicting daily risk of CDI by tailoring the algorithms to 2 individual institutional datasets. These AUROCs demonstrate the potential value of site-specific customization of CDI MLA.[40]

All architectures evaluated in our study demonstrated excellent discrimination (AUROC > 0.80) on the hold-out test dataset. In contrast to the modest performance of neural networks reported by Marra *et al*.,[45] the 1D-CNN and D-LSTM models in our study achieved high discrimination, representing the first report to the authors' knowledge of such high performance by a neural network for a CDI prediction task.

This research also has several limitations. Given the retrospective nature of this study, we were unable to determine the performance of the MLAs in a prospective clinical setting, or in settings where data availability and data collection frequency may differ from the hospitals included in the development and external datasets. Prospective validation is required to evaluate clinician response to CDI risk predictions, and whether such predictions may significantly improve metrics of CDI patient outcomes, hospital infection control and CDI-associated cost burden. Such evaluation may be complemented by concurrent studies on factors motivating or inhibiting clinician acceptance of MLAs to predict CDI, the feasibility of using MLA design elements, and educational initiatives to enhance the acceptability and adoptability of these tools. The gold standard does not distinguish between community-acquired and hospital-associated CDI, which may be a relevant distinction for future iterations of CDI MLAs, given the greater severity and higher mortality associated with hospital-associated CDI relative to community-acquired CDI.[47] Lastly, our gold standard included all diagnostic tests currently recommended by IDSA, including nucleic acid amplification tests and toxin enzyme immunoassays. These tests alone cannot discriminate between symptomatic CDI and asymptomatic colonization.[13] However, the use of a clinical gold standard as the training gold standard represents a strength of this study, which may be used as the foundation for future research incorporating additional symptom data to further enhance the precision of the prediction outcome.

CONCLUSIONS

We have demonstrated that MLAs using just the first 6 hours of hospitalization data can predict CDI with high discrimination. We have also shown that XGBoost can achieve comparable predictive performance to the more complex neural networks, and that different training techniques to account for the low prevalence of CDI in training data are optimal for different MLA architectures. Future research may build upon this work by validating MLA-based CDI prediction tools on prospectively collected, live data, and soliciting feedback from clinician target end users to optimize the usefulness and acceptability of MLA alerts of CDI risk.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.ajic.2021.11.012.

References

1. Riddle DJ, Dubberke ER. Trends in clostridium difficile disease: epidemiology and intervention. *Infect. Med*. 2009;26:211–220.
2. Collins CE, Ayturk MD, Flahive JM, Emhoff TA, Anderson FA, Santry HP. Epidemiology and outcomes of community-acquired clostridium difficile infections in medicare beneficiaries. *J. Am. Coll. Surg*. 2014;218:1141–1147. e1.
3. Zimlichman E, Henderson D, Tamir O, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA Intern. Med*. 2013;173:2039–2046.
4. Guh AY, Mu Y, Winston LG, et al. Trends in U.S. burden of clostridioides difficile infection and outcomes. *N. Engl. J. Med*. 2020;382:1320–1330.
5. Lessa FC, Mu Y, Bamberg WM, et al. Burden of clostridium difficile infection in the United States. *N. Engl. J. Med*. 2015;372:825–834.
6. Czepiel J, Dróżdż M, Pituch H, et al. Clostridium difficile infection: review. *Eur. J. Clin. Microbiol. Infect. Dis*. 2019;38:1211–1221.
7. Jodlowski TZ, Oehler R, Kam LW. Melnychuk, I. emerging therapies in the treatment of clostridium difficile-associated disease. *Ann. Pharmacother*. 2006; 40:2164–2169.
8. Kyne L, Hamel MB, Polavaram R, Kelly CP. Health care costs and mortality associated with nosocomial diarrhea due to clostridium difficile. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am*. 2002;34:346–353.
9. Balsells E, Filipescu T, Kyaw MH, Wiuff C, Campbell H, Nair H. Infection Prevention and Control of Clostridium Difficile: A Global Review of Guidelines, Strategies, and Recommendations. *J. Glob. Health*. 2016;6: 020410.
10. Exemptions (2018 Requirements). Accessed April, 21, 2021. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/common-rule-subpart-a-46104/index.html.
11. Dial S, Delaney JaC, Barkun AN, Suissa S. Use of gastric acid-suppressive agents and the risk of community-acquired clostridium difficile-associated disease. *JAMA*. 2005;294:2989–2995.
12. Eze P, Balsells E, Kyaw MH, Nair H. Risk factors for clostridium difficile infections – an overview of the evidence base and challenges in data synthesis. *J. Glob. Health*. 2021;7:010417.
13. McDonald LC, Gerding DN, Johnson S, et al. Clinical Practice Guidelines for Clostridium Difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am*. 2018;66:e1–e48.
14. Radhachandran A, Garikipati A, Iqbal Z, et al. A machine learning approach to predicting risk of myelodysplastic syndrome. *Leuk. Res*. 2021;109:106639.
15. Giang C, Calvert J, Rahmani K, et al. Predicting ventilator-associated pneumonia with machine learning. *Medicine (Baltimore)*. 2021;100:e26246.
16. Ryan L, Mataraso S, Siefkas A, et al. A machine learning approach to predict deep venous thrombosis among hospitalized patients. *Clin. Appl. Thromb. Off. J. Int. Acad. Clin. Appl. Thromb*. 2021;27: 1076029621991185.
17. A Novel Deep Similarity Learning Approach to Electronic Health Records Data. *IEEE Journals & Magazine | IEEE Xplore*. 2021.. https://ieeexplore.ieee.org/document/9257424. accessed 2021-06-28.
18. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. *J. Biomed. Inform*. 2017;69:218–229.
19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016785–794. ACM: San Francisco California USA.
20. Marra AR, Perencevich EN, Nelson RE, et al. Incidence and outcomes associated with clostridium difficile infections: a systematic review and meta-analysis. *JAMA Netw. Open*. 2020;3: e1917597.
21. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng*. 2009;21:1263–1284.
22. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv*. 2016;49. 31:1-31:50.
23. A Recipe for Training Neural Networks. 2021. Accessed May 5, 2021. http://karpathy.github.io/2019/04/25/recipe/#2-set-up-the-end-to-end-trainingevaluation-skeleton–get-dumb-baselines.
24. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *J. Med. Chem*. 2020;63:8761–8777.
25. Tabak YP, Johannes RS, Sun X, Nunez CM, McDonald LC. Predicting the Risk for Hospital-Onset Clostridium Difficile Infection (HO-CDI) at the Time of Inpatient Admission: HO-CDI risk score. *Infect. Control Hosp. Epidemiol*. 2015;36:695–701.
26. Dubberke ER, Yan Y, Reske KA, et al. Development and validation of a clostridium difficile infection risk prediction model. *Infect. Control Hosp. Epidemiol*. 2011;32:360–366.
27. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol*. 2010;5:1315–1316.
28. Swift A, Heale R, Twycross A. What are sensitivity and specificity? *Evid. Based Nurs*. 2020;23:2–4.
29. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. *Data Min. Knowl. Discov*. 2019;33:917–963.
30. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf. Proc*. 2016;56:301–318.
31. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. *Machine Learning for Healthcare Conference*. 201673–100. PMLR.
32. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digit. Med*. 2018;1:1–10.
33. Chang D, Chang D, Pourhomayoun M. Risk prediction of critical vital signs for ICU patients using recurrent neural network. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. 20191003–1006.
34. Luz CF, Vollmer M, Decruyenaere J, Nijsten MW, Glasner C, Sinha B. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. *Clin. Microbiol. Infect*. 2020;26:1291–1299.
35. Lv H, Yang X, Wang B, et al. machine learning–driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: retrospective study. *J. Med. Internet Res*. 2021;23:e24996.
36. Lever J, Krzywinski M, Altman N. Model Selection and Overfitting. *Nat. Methods*. 2016;13:703–704.
37. Kassam Z, Cribb Fabersunne C, Smith MB, et al. Clostridium Difficile Associated Risk of Death Score (CARDS): a novel severity score to predict mortality among

hospitalised patients with C. Difficile infection. *Aliment. Pharmacol. Ther.* 2016;43:725–733.
38. Miller MA, Louie T, Mullane K, et al. Derivation and Validation of a Simple Clinical Bedside Score (ATLAS) for clostridium difficile infection which predicts response to therapy. *BMC Infect. Dis.* 2013;13:148.
39. Arora V, Kachroo S, Ghantoji SS, Dupont HL, Garey KW. High horn's index score predicts poor outcomes in patients with clostridium difficile infection. *J. Hosp. Infect.* 2011;79:23–26.
40. Oh J, Makar M, Fusco C, et al. Generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *Infect. Control Hosp. Epidemiol.* 2018;39:425–433.
41. Escobar GJ, Baker JM, Kipnis P, et al. Prediction of recurrent clostridium difficile infection using comprehensive electronic medical records in an integrated healthcare delivery system. *Infect. Control Hosp. Epidemiol.* 2017;38:1196–1203.
42. Steele S, Bilchik A, Eberhardt J, et al. Using machine-learned bayesian belief networks to predict perioperative risk of clostridium difficile infection following colon surgery. interact. *J. Med. Res.* 2012;1:e2131.
43. Wiens J, Campbell WN, Franklin ES, Guttag JV, Horvitz E. Learning data-driven patient risk stratification models for clostridium difficile. *Open Forum Infect. Dis.* 2014;1:ofu045.
44. Kuntz JL, Smith DH, Petrik AF, et al. Predicting the risk of clostridium difficile infection upon admission: a score to identify patients for antimicrobial stewardship efforts. *Perm. J.* 2016;20:20–25.
45. Marra AR, Alzunitan M, Abosi O, et al. Modest Clostridiodes Difficile Infection Prediction Using Machine Learning Models in a Tertiary Care Hospital. *Diagn. Microbiol. Infect. Dis.* 2020;98: 115104.
46. Li BY, Oh J, Young VB, Rao K, Wiens J. Using machine learning and the electronic health record to predict complicated clostridium difficile infection. *Open Forum Infect. Dis.* 2019;6:ofz186.
47. Khanna S, Pardi DS, Aronson SL, et al. The epidemiology of community-acquired clostridium difficile infection: a population-based study. *Am. J. Gastroenterol.* 2012;107:89–95.