Open Forum Infectious Diseases

MAJOR ARTICLE

A Predictive Model to Identify Complicated *Clostridiodes difficile* Infection

Jeffrey A. Berinstein,[1,a] Calen A. Steiner,[2,3,a] Samara Rifkin,[1,a] D. Alexander Perry,[4] Dejan Micic,[5] Daniel Shirley,[6] Peter D. R. Higgins,[1] Vincent B. Young,[7,8] Allen Lee,[1,b,©] and Krishna Rao[7,b]

[1]Division of Gastroenterology and Hepatology, University of Michigan, Ann Arbor, Michigan, USA, [2]Division of Gastroenterology and Hepatology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA, [3]Department of Medicine and the Mucosal Inflammation Program, University of Colorado School of Medicine, Aurora, Colorado, USA, [4]Division of Infectious Diseases, University of Arizona, Tucson, Arizona, USA, [5]Section of Gastroenterology, Hepatology, and Nutrition, University of Chicago, Chicago, Illinois, USA, [6]Division of Infectious Diseases, University of Wisconsin, Madison, Wisconsin, USA, [7]Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan, USA, and [8]Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan, USA

Background. *Clostridioides difficile* infection (CDI) is a leading cause of health care–associated infection and may result in organ dysfunction, colectomy, and death. Published risk scores to predict severe complications from CDI demonstrate poor performance upon external validation. We hypothesized that building and validating a model using geographically and temporally distinct cohorts would more accurately predict risk for complications from CDI.

Methods. We conducted a multicenter retrospective cohort study of adults diagnosed with CDI. After randomly partitioning the data into training and validation sets, we developed and compared 3 machine learning algorithms (lasso regression, random forest, stacked ensemble) with 10-fold cross-validation to predict disease-related complications (intensive care unit admission, colectomy, or death attributable to CDI) within 30 days of diagnosis. Model performance was assessed using the area under the receiver operating curve (AUC).

Results. A total of 3646 patients with CDI were included, of whom 217 (6%) had complications. All 3 models performed well (AUC, 0.88–0.89). Variables of importance were similar across models, including albumin, bicarbonate, change in creatinine, non-CDI-related intensive care unit admission, and concomitant non-CDI antibiotics. Sensitivity analyses indicated that model performance was robust even when varying derivation cohort inclusion and CDI testing approach. However, race was an important modifier, with models showing worse performance in non-White patients.

Conclusions. Using a large heterogeneous population of patients, we developed and validated a prediction model that estimates risk for complications from CDI with good accuracy. Future studies should aim to reduce the disparity in model accuracy between White and non-White patients and to improve performance overall.

Keywords. *Clostridiodes difficile*; machine learning; predictive model; severe disease.

Clostridiodes difficile infection (CDI) is the leading cause of health care–associated infection in US hospitals, accounting for nearly 500 000 infections per year [1, 2]. Furthermore, ~8% of patients with CDI may develop disease-related complications, including organ dysfunction, severe sepsis, colectomy, and death [3]. In addition to the substantial patient morbidity and mortality, CDI results in $4.8 billion in acute health care costs in the United States, with even more costs associated with non–acute care settings [4]. While treatments to reduce the risk of adverse outcomes exist, it is not optimal to use them in all patients due to cost, invasive nature, and/or experimental status (eg, fecal transplant or colectomy). Thus, there is a need to identify patients at risk for adverse outcomes from CDI.

Although risk scores to identify patients at risk for CDI-related complications have been developed, these scoring systems have limited generalizability due to being developed in small, single-center cohorts and have not undergone external validation [5–13]. Recently, our group demonstrated that published CDI severity scoring systems performed poorly when tested on a large, multicenter cohort within the United States [14]. Thus, an accurate prediction model that can be applied early after CDI diagnosis is needed to identify patients at risk for complications from CDI, which may allow for more equitable allocation of treatments to minimize adverse CDI outcomes. In this study, we aimed to determine whether using structured electronic health record data from several geographically distinct centers in the United States would provide a more generalizable predictive model for complicated CDI.

Received 06 October 2022; editorial decision 24 January 2023; accepted 31 January 2023; published online 2 February 2023

[a]Co-first authors, equal contribution

[b]Co-senior authors, equal contribution

Correspondence: Allen Lee, MD, MS, 3912 Taubman Center, 1500 E. Medical Center Drive, SPC 5362, Ann Arbor, MI 48109 (allenlee@umich.edu); or Krishna Rao, MD, MS, 1150 West Medical Center Drive, MSRB I, Room 1510B, Ann Arbor, MI, 48109 (krirao@med.umich.edu).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofad049

METHODS

Patient Cohorts

We conducted a retrospective cohort study at 4 geographically and temporally distinct cohorts, including (1) the University of Michigan (UM) from 2010 to 2012 and (2) 2015 to 2016; (3) University of Wisconsin (UW) from 2014 to 2015; and (4) the University of Chicago (UC) from 2013 to 2015, as previously described [14]. Adult subjects age ≥18 years diagnosed with CDI were included in our analysis. CDI was diagnosed by the presence of diarrhea (≥3 unformed stools in a 24-hour period) and positive stool testing. At UW and UC, a diagnosis of CDI was made using a positive real-time polymerase chain reaction (PCR) for the *tcdB* gene (Simplexa *C. difficile* Universal Direct, Diasorin Molecular LLC, Cypress, CA, USA). At UM, a diagnosis of CDI was made using 2-step mechanism testing in which *C. difficile* glutamate dehydrogenase (GDH) and toxins A or B (C. Diff Quik Chek Complete, Alere, Waltham, MA, USA) were evaluated, and discordant result (GDH+/toxin−) stool tests were subjected to analysis for the tcdB gene by real-time PCR.

Primary Outcome

We defined CDI as complicated if it led to any of 3 adverse outcomes within 30 days of CDI diagnosis: admission to intensive care unit (ICU), colectomy, or death attributable to CDI as determined by the study team physicians [12, 15]. If patients had a positive CDI diagnosis but were admitted to the ICU, underwent colectomy, or died for reasons not attributable to CDI, they were not included in our composite outcome for complicated CDI. For example, a patient admitted to the ICU for septic shock secondary to pneumonia while actively being treated for CDI would not have met our primary outcome definition. Complicated CDI was determined manually by chart review at each center by an infectious disease specialist or gastroenterologist with expertise in *C. difficile*. A diagnosis of complicated CDI could be made independently of when the CDI was confirmed as attributions were made in a retrospective manner (after discharge) when the complete hospital course was available for review. Clinical and demographic variables including comorbidities, medications, vitals, laboratory results, and study results (such as radiographic imaging) were collected for each patient's admission through automated query of the electronic health record (EHR).

Patient Consent

Written informed consent was not obtained as this study did not include factors necessitating patient consent. The institutional review boards at UM, UW, and UC gave ethical approval for this work.

Predictor Variables

A total of 32 predictor variables were evaluated for inclusion in the final model based on literature review of clinically relevant factors (Supplementary Table 1). All variables had to be collected within 48 hours of CDI diagnosis to be included for analysis. Non-CDI-related ICU admission and non-CDI concurrent antibiotics were included as predictor variables only if ICU admission and/or antibiotic use were unrelated to CDI. For example, a patient admitted to the ICU postoperatively who subsequently developed CDI would not be eligible to meet criteria for the composite outcome (ie, ICU, colectomy, death), but non-CDI-related ICU admission would be included as a predictor of CDI-related complication.

Data Preprocessing

We used R, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria), for cleaning and munging data. Data were randomly split, with 75% of the data used for model training and the remaining 25% data held out for testing and validation. The data were stratified by the proportion of severe CDI events so that the distribution of the outcome was maintained in both the training set and the test set. We intentionally did not split the data by cohort in order to maintain the generalizability of our results across centers and time. However, we did investigate site-specific variations in patient demographic composition and clinical practice patterns around the diagnosis and management of CDI (see sensitivity analysis below). Missing data were imputed using the R package missForest [16], a random forest–based multiple imputation method previously shown to have the lowest imputation error for both continuous and categorical variables [17]. Of note, only laboratory data and vital signs were imputed, while demographic data, such as age, gender, and race, and comorbid conditions were not imputed as there were no missing data for these variables. Numeric variables were centered and scaled while categorical variables were recoded into dummy variables.

Model Development and Testing

With the randomly selected 75% of subjects included in the training set, we developed 3 separate machine learning classification models using the *Tidymodels* framework [18], including least absolute shrinkage and selection operator (lasso) L1-regularized regression with the R package *glmnet* [19], random forest with the R package *ranger* [20], and extreme gradient–boosted trees with the R package *XGBoost* [21]. The best performing algorithms were combined in an ensemble model using stacking with the R package *stacks* [22]. Machine learning algorithms were first applied to training data to parameterize and fit the model. Ten-fold cross-validation was utilized to estimate model accuracy and tune model hyperparameters. To evaluate the prediction accuracy of machine learning models, areas under the receiver operating characteristic curve (AUCs) were calculated for each model using the independent test data consisting of the remaining 25% of patients not selected for the training set (Figure 1).

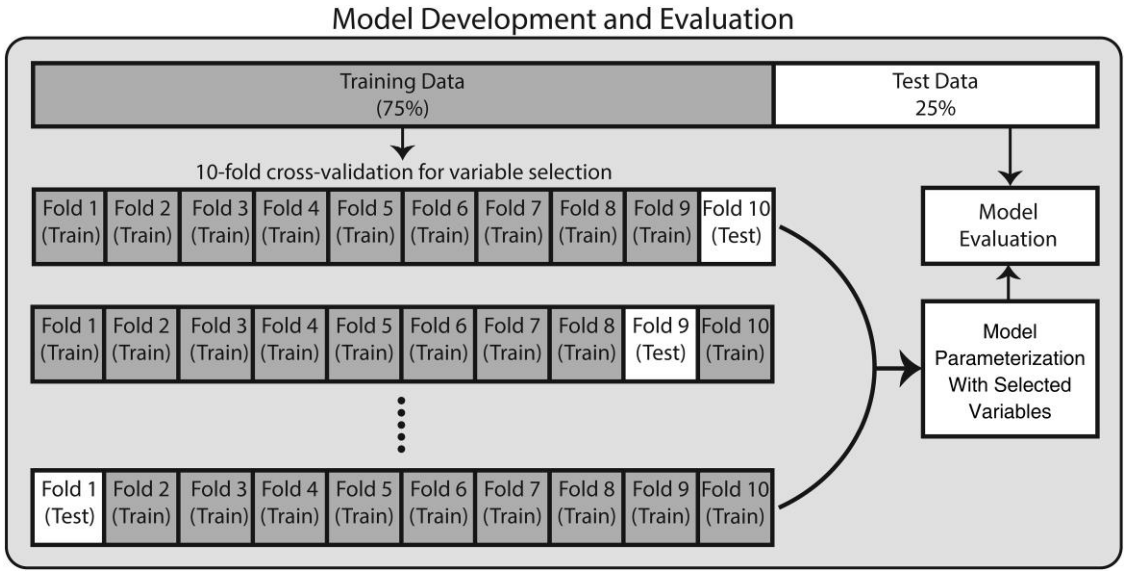

Figure 1. Model development and validation strategies. Data from University of Michigan 2010–2012 and 2015–2016, University of Wisconsin 2014–2015, and University of Chicago 2013–2015 were randomly split into a training set (75%) and an independent test set (25%). Four separate classification models, including least absolute shrinkage and selection operator (lasso), random forest, extreme gradient boosted trees (XGBoost), and ensemble model using stacking were evaluated. Ten-fold cross-validation was utilized to tune model hyperparameters. Model accuracy was then evaluated by calculating area under the receiver operating characteristic curves and calibration curves for each model using independent test data. Variable importance and breakdown plots were also performed to determine feature importance for each model.

Variable Importance

Variable importance was determined by using the R package *vip* [23], which provides model-specific variable importance scores. To further improve interpretability, we also performed locally interpretable model-agnostic explanations using the R package *breakDown* [24], which decomposes model predictions into parts that can be attributed to particular variables.

Sensitivity Analyses

We performed several post hoc analyses to assess model performance and generalizability with regards to differences in study sites, time periods, clinical practice patterns around the diagnosis and management of CDI, and participant demographic composition.

RESULTS

Patient Characteristics

A total of 3646 patients who were diagnosed with CDI were analyzed from 4 cohorts (Table 1). The total population had a mean age (SD) of 58.2 (18.0) years, with 53.1% of the population being female. Age and sex were similar among the 4 cohorts. UC was composed of a Black majority (53.2%), whereas all other sites were predominantly White (81.6%–92.2%). In addition, the UM 2015–2016 cohort was the only cohort with outpatients (27.4%). A total of 217 patients (6.0%) met the primary end point, including 64 (4.8%) in the UC cohort; 90 (7.9%) in the UM 2010–2012 cohort; 28 (4.3%) in the UM 2015–2016 cohort; and 35 (6.8%) in the UW cohort. Mean laboratory values, vital signs, and comorbid conditions are presented by cohort in Table 1.

Model Training and Performance

Lasso regression, random forest, and stacked ensemble models all performed well, with AUC scores ranging from 0.88 to 0.89 (Figure 2; Supplementary Table 2) when tested on an independent test data set. XGBoost models performed poorly (AUC, 0.50) and were not carried forward in subsequent analyses. Model calibration plots for the 3 models are illustrated in Supplementary Figure 1. The calibration plots for the 3 models show minimal mismatch between the probabilities predicted by the model and the probabilities observed in the data, with the stacked ensemble model demonstrating the best calibration (closest to the straight diagonal reference line).

Sensitivity Analyses

Generalizability of Models

For the site-specific cohort analysis, we trained the machine learning algorithms on 3 cohorts and validated model performance on the fourth cohort. We repeated this process 3 times so that models were built on the other 3 and then validated on each individual remaining cohort (Supplementary Figure 2). Because our cohorts included subjects diagnosed from 2012 to 2016, this process also served as a temporal validation. We found that models trained on only 3 cohorts were able to predict complicated CDI in the fourth cohort with high accuracy. Model performance was robust in this sensitivity analysis

Table 1. Patient Characteristics

	University of Chicago (2013–2015) (n = 1341)	University of Michigan (2010–2012) (n = 1144)	University of Michigan (2015–2016) (n = 646)	University of Wisconsin (2014–2015) (n = 515)	Total Patient Population (n = 3646)
Age, mean (SD), y	58.7 (18.5)	57.3 (18.0)	57.7 (18.2)	59.2 (16.1)	58.2 (18.0)
Female, No. (%)	703 (52.4)	625 (54.6)	344 (53.3)	264 (51.3)	1936 (53.1)
Race, No. (%)					
White	545 (40.6)	933 (81.6)	558 (86.4)	475 (92.2)	2511 (68.9)
Black	713 (53.2)	147 (12.8)	46 (7.1)	26 (5.0)	932 (25.6)
Other	83 (6.2)	64 (5.6)	42 (6.5)	14 (2.7)	203 (5.6)
Inpatient admission, No. (%)	1341 (100.0)	1144 (100.0)	469 (72.6)	515 (100.0)	3469 (95.1)
Non-CDI ICU admission, No. (%)	84 (6.3)	144 (12.6)	9 (1.4)	61 (11.8)	298 (8.2)
Disease-related complications from *C. difficile*, No. (%)					
30-d mortality	39 (2.9)	49 (4.3)	23 (3.6)	17 (3.3)	128 (3.5)
30-d colectomy	16 (1.2)	4 (0.3)	1 (0.2)	5 (1.0)	26 (0.7)
30-d ICU admission	18 (1.3)	49 (4.3)	5 (0.8)	26 (5.0)	98 (2.7)
Total composite end point, mean (SD)	64 (4.8)	90 (7.9)	28 (4.3)	35 (6.8)	217 (6.0)
Concomitant non-CDI antibiotic use within 30 d, No. (%)	908 (67.7)	756 (66.1)	228 (35.3)	354 (68.7)	2246 (61.6)
Peak WBC count, mean (SD), K/μL	11.3 (12.0)	13.4 (12.4)	12.2 (15.5)	13.7 (20.2)	12.5 (14.1)
Baseline creatinine, mean (SD), mg/dL	1.6 (2.2)	1.4 (1.7)	1.2 (1.3)	1.5 (1.9)	1.5 (1.9)
Peak creatinine, mean (SD), mg/dL	2.1 (2.4)	1.6 (1.8)	1.3 (1.4)	2.0 (2.4)	1.8 (2.1)
Creatinine change, mean (SD), mg/dL	0.5 (1.1)	0.3 (1.1)	0.1 (0.7)	0.5 (1.0)	0.4 (1.0)
Acute kidney injury, No. (%)					
None	734 (54.7)	797 (69.7)	299 (46.3)	289 (56.1)	2119 (58.1)
Stage 1	588 (43.8)	336 (29.4)	47 (7.3)	159 (30.9)	1130 (31.0)
Stage 2	3 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	3 (0.1)
Stage 3	734 (54.7)	797 (69.7)	299 (46.3)	289 (56.1)	2119 (58.1)
Lowest albumin, mean (SD), g/dL	3.034 (0.707)	3.157 (0.664)	3.305 (0.739)	2.650 (0.676)	3.083 (0.716)
Lowest hemoglobin, mean (SD), g/dL	9.725 (1.889)	9.489 (1.990)	10.016 (2.346)	9.253 (2.239)	9.628 (2.062)
Peak platelets, mean (SD), K/μL	250.434 (147.188)	258.749 (186.944)	259.821 (135.329)	217.984 (125.550)	250.003 (157.399)
Lowest sodium, mean (SD), mmol/L	136.008 (4.728)	136.508 (4.214)	137.767 (3.790)	136.903 (4.388)	136.571 (4.419)
Lowest bicarbonate, mean (SD), mmol/L	20.090 (4.524)	23.603 (4.527)	24.403 (4.412)	22.606 (4.453)	21.611 (4.836)
Maximum body temperature, mean (SD), °F	98.924 (0.785)	99.526 (1.462)	99.982 (1.616)	99.759 (1.499)	99.395 (1.340)
Maximum systolic blood pressure, mean (SD), mmhg	105.965 (9.102)	99.329 (19.353)	93.894 (18.664)	99.479 (16.855)	101.164 (16.142)
No. of positive prior CDIs, mean (SD)	0.192 (0.394)	0.266 (0.442)	0.289 (0.454)	0.058 (0.234)	0.213 (0.410)
Peripheral vascular disease, No. (%)	109 (8.1)	75 (6.6)	153 (23.7)	107 (20.8)	444 (12.2)
Peptic ulcer disease, No. (%)	0 (0.0)	27 (2.4)	67 (10.4)	4 (0.8)	98 (2.7)
Congestive heart failure, No. (%)	323 (24.1)	151 (13.2)	153 (23.7)	129 (25.0)	756 (20.7)
Malignancy, No. (%)	225 (16.8)	220 (19.2)	0 (0.0)	515 (100.0)	960 (26.3)
Metastatic malignancy, No. (%)	156 (11.6)	61 (5.3)	0 (0.0)	41 (8.0)	258 (7.1)

Table 1. Continued

	University of Chicago (2013–2015) (n = 1341)	University of Michigan (2010–2012) (n = 1144)	University of Michigan (2015–2016) (n = 646)	University of Wisconsin (2014–2015) (n = 515)	Total Patient Population (n = 3646)
Chronic pulmonary disease, No. (%)	333 (24.8)	320 (28.0)	205 (31.7)	157 (30.5)	1015 (27.8)
Rheumatoid arthritis, No. (%)	189 (14.1)	77 (6.7)	39 (6.0)	39 (7.6)	344 (9.4)
Diabetes without complication, No. (%)	290 (21.6)	267 (23.3)	187 (28.9)	171 (33.2)	915 (25.1)
Diabetes with complication, No. (%)	100 (7.5)	127 (11.1)	0 (0.0)	94 (18.3)	321 (8.8)
Renal disease, No. (%)	410 (30.6)	309 (27.0)	230 (58.2)	196 (38.1)	1145 (33.7)
Obesity, No. (%)	116 (8.7)	72 (6.3)	164 (25.9)	18 (3.5)	370 (10.2)
Inflammatory bowel disease, No. (%)	97 (7.2)	110 (9.6)	108 (16.7)	26 (5.0)	231 (9.2)

Abbreviations: CDI, *Clostridioides difficile* infection; ICU, intensive care unit; WBC, white blood cell.

(AUC ranging from 0.84 to 0.92), although there was a drop in performance when the data from UC were used as the test set (AUC, 0.75–0.76) (Figure 3; Supplementary Table 2).

Impact of Non-CDI-Related ICU Admission as Predictor Variable

Given the potential for shared information between CDI-related complications as an outcome and non-CDI-related ICU admission, we performed a sensitivity analysis to determine whether model performance was affected when excluding non-CDI-related ICU admission as a predictor variable (lasso model: AUC, 0.82; 95% CI, 0.77–0.88; random forest: AUC, 0.82; 95% CI, 0.77–0.89; stacked ensemble: AUC, 0.83; 95% CI, 0.77–0.89) (Supplementary Figure 3, Supplementary Table 2).

PCR vs Two-Step Diagnosis for CDI

As the definition of CDI varied by individual centers, we performed an additional sensitivity analysis to determine whether diagnosis of CDI by PCR only vs 2-step mechanism (PCR screen with enzyme immunoassay confirmation) influenced our model predictions. Sites that utilized PCR only (ie, UW and UC) were analyzed separately from sites that utilized a 2-step diagnostic approach (ie, UM 2010 and 2016). We again randomly split the data, training the model on 75% of the data, and tested and validated our models on the remaining held-out 25% of the data (Supplementary Figure 4). For sites that used 2-step testing (ie, UM 2010 and 2016), our models retained excellent performance (AUC, 0.89–0.91), while model performance was lower when using sites that used PCR testing alone (ie, UW and UC) as the hold-out test set (AUC, 0.79–0.84) (Supplementary Figure 5, Supplementary Table 2).

Effect of Race on Model Prediction

Finally, we noted our discrepant model performance in the UC cohort, which included a larger proportion of subjects self-reporting as Black race. Given this observation and the growing concern about implicit bias with machine learning algorithms [25, 26], we performed a fourth set of sensitivity analyses to determine whether self-reported race may affect model predictions. We first determined whether machine learning algorithms may predict race. Because of low numbers of Hispanic/Latino, Asian, American Indian/Alaska Native, or mixed races at each site, patients were grouped into White vs non-White categories. Second, as UC had a much higher proportion of patients self-reporting as Black compared with other sites, we trained and validated models using data from White patients only at UC and compared the prediction with a model trained and validated on the entire cohort from UC. Third, using data from all 4 cohorts, we stratified our models by race to determine if race affected model prediction (Supplementary Figure 6). Using data from all 4 cohorts, we first determined that lasso and random forest algorithms were able to predict race with reasonable accuracy (AUC, 0.75 for lasso; AUC, 0.78 for random forest).

Next, using data only from White patients at UC, our models showed good performance for predicting disease-related complications from CDI (AUC, 0.87 for both lasso and random forest). However, we found that performance worsened when we trained and validated the model on all patients at UC (AUC, 0.76 for lasso; AUC, 0.75 for random forest).

We then investigated whether race was predictive of complicated CDI. We found that non-White race was not associated with complicated CDI in all cohorts ($P = .71$) as well as in each cohort ($P = .53$, $P = .15$, $P = .46$, and $P = .68$ for UC, UW, UM 2010, and 2016, respectively). We further found that non-White race was not associated with individual components of the composite outcome (ie, ICU admission, colectomy, or death related to CDI) for the entire cohort as well as for each individual cohort (data not shown).

Finally, when using data from only White patients in all 4 cohorts, our models showed good performance for predicting complicated CDI (AUC, 0.84 for lasso; AUC, 0.86 for random forest and stacked ensemble). However, when using data from non-White patients at all 4 cohorts, models showed worse

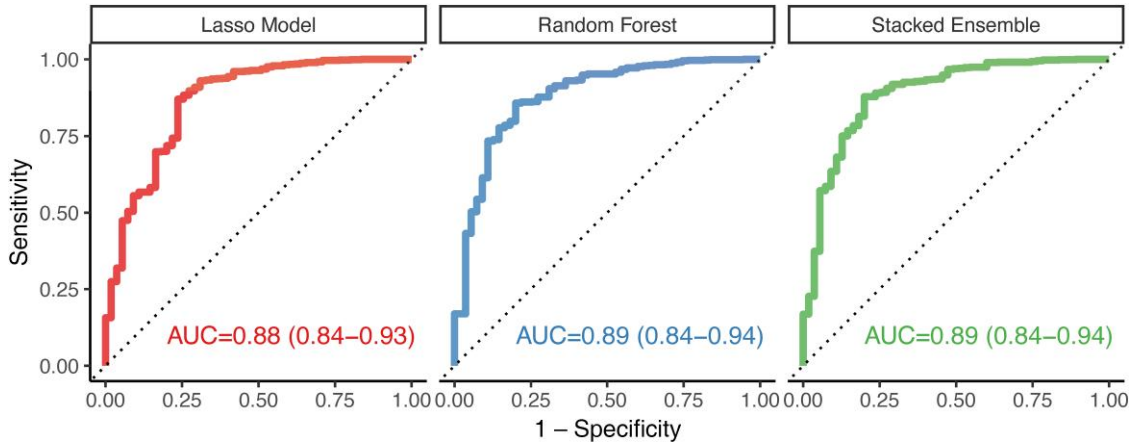

Figure 2. Receiver operator characteristic curves for lasso, random forest, and stacked ensemble models. After randomly splitting the data into training/validation sets, lasso regression, random forest, and stacked ensemble models were trained. All models demonstrated good performance when tested on an independent validation set (lasso regression: AUC, 0.88; 95% CI, 0.84–0.93; random forest: AUC, 0.89; 95% CI, 0.84–0.94; stacked ensemble: AUC, 0.88; 95% CI, 0.83–0.94). Abbreviation: AUC, area under the curve.

performance (AUC, 0.62 for lasso; AUC, 0.68 for random forest; AUC, 0.67 for stacked ensemble) (Supplementary Figure 7). Similar variables were found to be important for predicting complicated CDI in White and non-White patients, including baseline and peak serum creatinine levels, use of non-CDI-related antibiotics, non-CDI-related ICU admission, low systolic blood pressure, and low serum bicarbonate levels (Supplementary Figure 8).

Predictive Features for Complicated CDI

Using variable importance analysis, the variables of importance were similar across models but varied in their relative contribution to each model (Figure 4*A*, *B*). The top predictors shared across all models included albumin, bicarbonate, change in creatinine, systolic blood pressure, non-CDI-related ICU admission, and concurrent non-CDI antibiotics.

Breakdown plots were also generated to determine how each variable contributed to a final prediction. Absence of factors such as non-CDI-related ICU admission, concurrent non-CDI antibiotics, low bicarbonate, low systolic blood pressure, and peak WBC count was associated with decreased risk for complicated CDI (Figure 4*C*, *D*). In contrast, non-CDI-related ICU admission, high WBC count, low systolic blood pressure, low albumin level, and concurrent non-CDI antibiotics were associated with increased risk for complicated CDI (Figure 4*E*, *F*).

DISCUSSION

In this multisite cohort study, machine learning models based on structured electronic health record data predicted disease-related complications from CDI with good accuracy. Lasso regression, random forest, and stacked ensemble machine learning methods demonstrated respectable performance in predicting severe complications from CDI (ie, AUC of 0.88–0.89). Importantly, we intentionally developed models without site-specific variables, and our results suggest that this model is generalizable across centers and time, which is critical when considering the heterogeneity in patient populations and practice patterns across the United States. However, self-reported race was a significant modifier of model performance, with models performing significantly worse in patients of non-White race. Finally, although there was some model-specific variability, the predictors most important in discriminating disease-related complications from CDI were similar between models.

While several CDI severity and complication scoring systems have been developed previously, they generally were developed from single centers and were not externally validated [5–13]. In the largest external validation to date, our group recently reported that these models yielded AUC scores below 0.70 [14]. A recent study in a similarly sized cohort in Virginia also showed poor performance of published models upon external validation [27]. Thus, current models cannot reliably predict risk for severe complications from CDI.

There are multiple notable strengths that make our approach more generalizable. First, our models were developed using data from 3 geographically distinct sites, which were comprised of a heterogenous population. Second, our models were robust despite differences in temporal trends and practice patterns across sites. For example, CDI was diagnosed by positive PCR test alone at UC and UW, while UM utilized a 2-step algorithm. The fact that model performance was lower when using sites that used PCR testing alone suggests that false classification of disease due asymptomatic CDI colonization may have affected model accuracy. Furthermore, UM started to use vancomycin as first-line treatment for CDI starting in

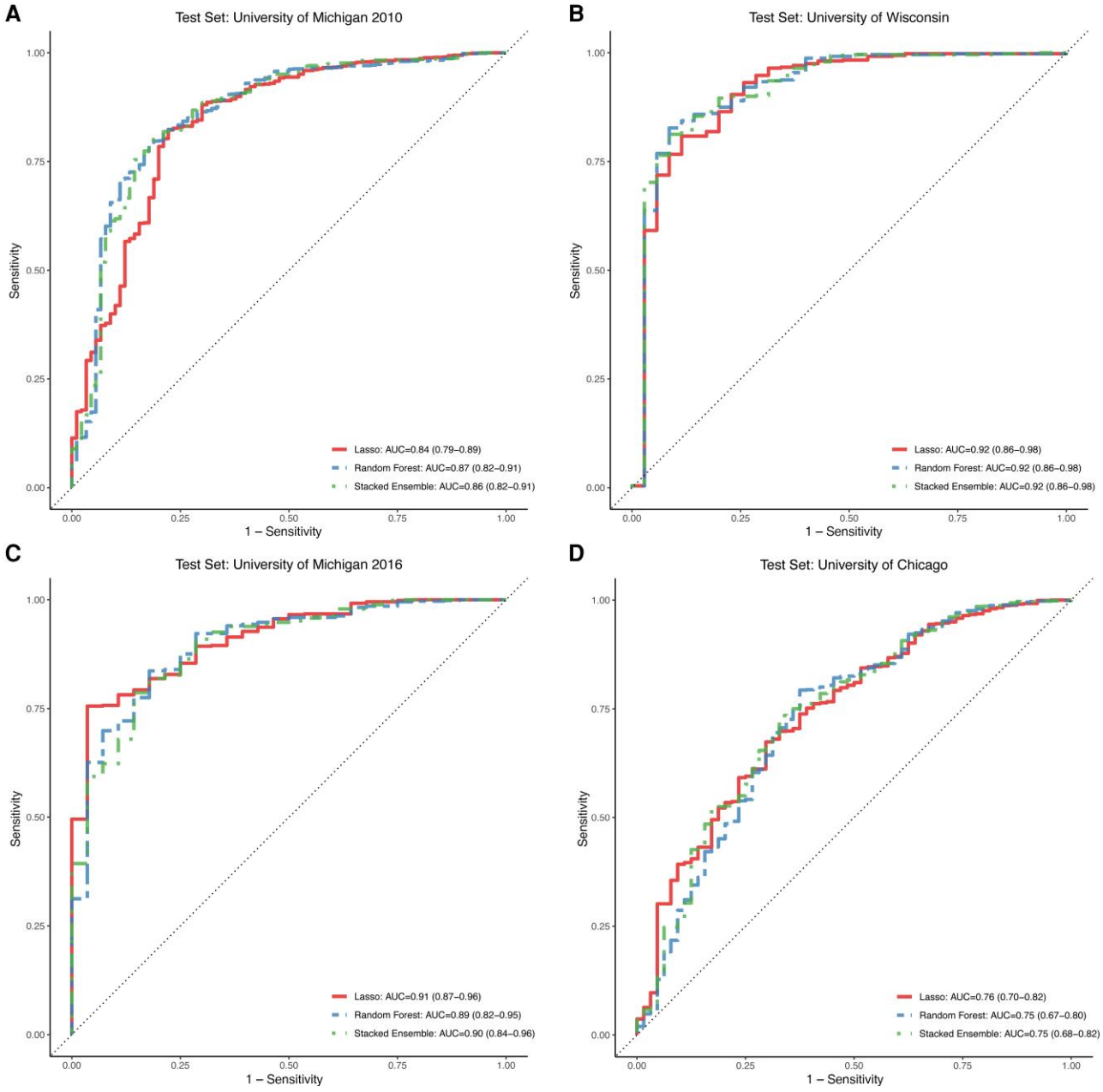

Figure 3. Models were robust despite geographical, demographic, and temporal variability. To determine whether models were generalizable across centers and across time, machine learning algorithms were derived and trained using data from 3 cohorts and then validated model performance on the fourth cohort (labeled "Test" in the figure). This process was repeated 3 times so that models were validated on each individual cohort. The lasso regression, random forest, and stacked ensemble models demonstrated good performance when tested on (*A*) the University of Michigan 2010–2012 cohort (lasso regression: AUC, 0.84; 95% CI, 0.79–0.89; random forest: AUC, 0.87; 95% CI, 0.82–0.91; stacked ensemble: AUC, 0.86; 95% CI, 0.82–0.91; (*B*) the University of Wisconsin cohort (lasso regression: AUC, 0.92; 95% CI, 0.86–0.98; random forest: AUC, 0.92; 95% CI, 0.86–0.98; stacked ensemble: AUC, 0.92; 95% CI, 0.86–0.98; and (*C*) the University of Michigan 2015–2016 cohort (lasso regression: AUC, 0.91; 95% CI, 0.87–0.96; random forest: AUC, 0.89; 95% CI, 0.82–0.95; stacked ensemble: AUC, 0.90; 95% CI, 0.84–0.96). Performance of the models dropped but still showed adequate performance when tested on (*D*) the University of Chicago cohort (lasso regression: AUC, 0.76; 95% CI, 0.70–0.82; random forest: AUC, 0.75; 95% CI, 0.67–0.80; stacked ensemble: AUC, 0.75; 95% CI, 0.68–0.82). Abbreviation: AUC, area under the curve.

2013, while the other centers did not switch from metronidazole to vancomycin as first-line treatment until mid-2016. Our sensitivity analyses demonstrated that our model retained good performance despite these differences in practice patterns and temporal trends.

While our model showed strong performance overall, we did see a noticeable drop in performance when the model was validated on the cohort from UC. As Black patients comprised a larger proportion of the cohort at UC compared with other sites, we speculated that this may be related to loss of important

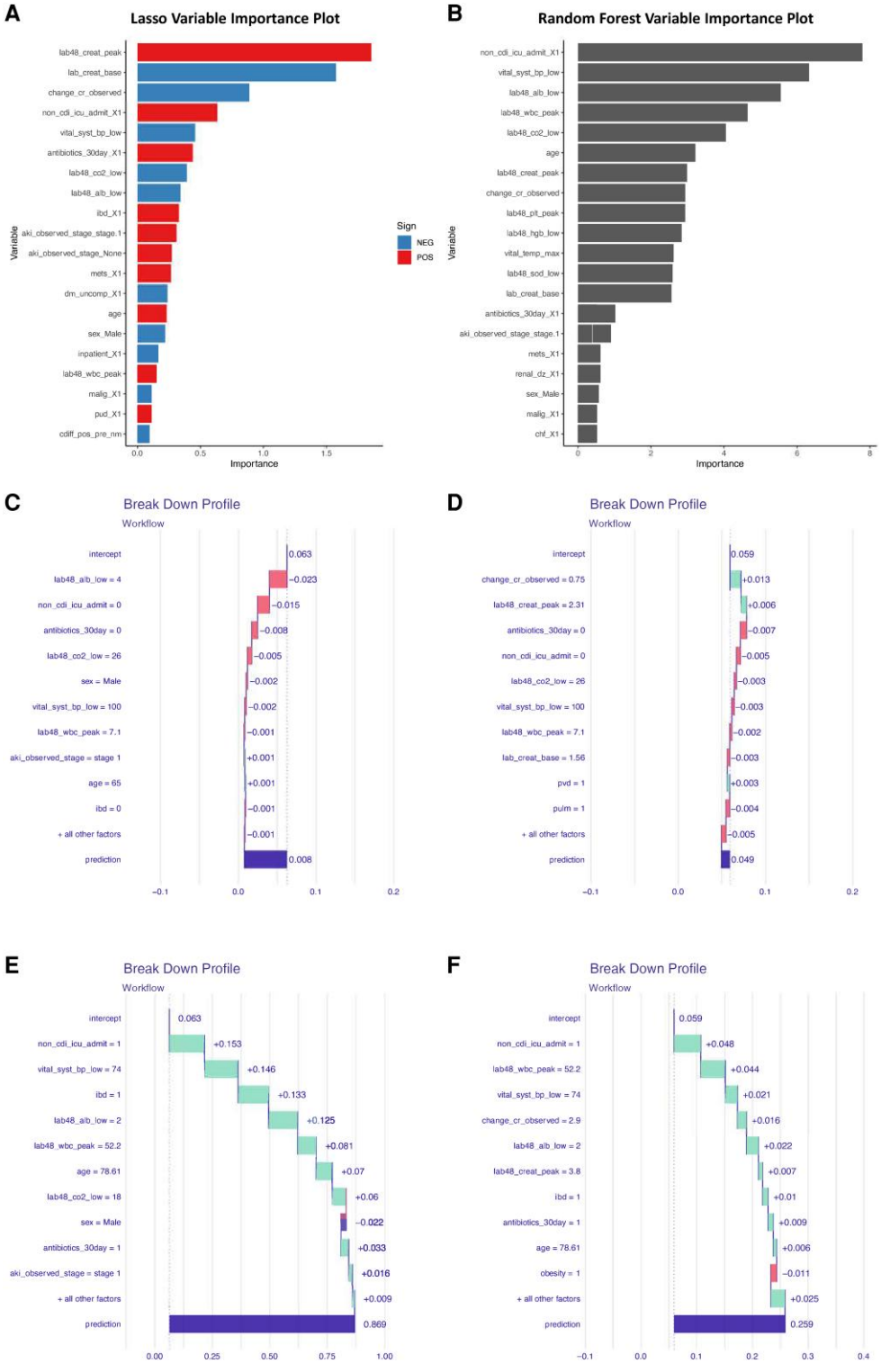

Figure 4. Predictive features for complicated *Clostridiodes difficile* infection. Twenty of the most important variables by variable importance analysis are illustrated for (*A*) lasso regression and (*B*) random forest. Please note that variable importance scores are assigned to each term in the model for linear models, such as lasso. Thus, variables positively associated with complicated CDI in the lasso model are shown in red, while variables negatively associated with complicated CDI are shown in blue. In contrast, absolute values are provided for nonlinear-based machine learning techniques, such as random forest, and no sign has been provided. Breakdown plots were also generated to determine how predictors vary for patients with low and high predicted probability for complicated CDI. Breakdown plots are shown for (*C*) lasso and (*D*) random forest models in the same patient (patient #1) with low predicted probability for complicated CDI. A distinct breakdown plot is depicted for (*E*) lasso and (*F*) random forest models in a patient with high predicted probability for complicated CDI (patient #277). The intercept represents the mean model-specific predicted probability for complicated CDI, while each subsequent variable increases or decreases predicted probability and results in the overall predicted probability (labeled prediction). Abbreviation: CDI, *Clostridiodes difficile* infection.

information related to racial diversity when model derivation was performed using data only from the UM or UW cohort. Indeed, our models retained excellent performance in White patients both at UC and from the entire cohort, while model performance in non-White patients was considerably less accurate. These results further raise the concern that machine learning algorithms may inadvertently result in systemic bias and health disparities [25, 26]. Interestingly, many of the most important variables for predicting disease-related complications from CDI when stratified by race involved serum creatinine levels, which are used to calculate estimated glomerular filtration rate (eGFR). As eGFR was initially developed using data from White patients [28], prior studies have demonstrated that non-Hispanic Black adults had higher serum creatinine levels compared with non-Hispanic White adults [29]. Thus, our models may have been underestimating eGFR in Black individuals, which may have biased our results. Furthermore, our results may be detecting real site-specific disparities in CDI complications related to race. However, as our models were developed to be site-agnostic, this may have limited our ability to adjust for these site-specific differences. Thus, these site-specific disparities should be identified and addressed in subsequent studies. Future research should also focus on whether including larger populations of non-White patients and carefully selecting nonbiased variables, for example, cystatin C–based estimates of GFR [30, 31], or other host- and/or microbe-derived biomarkers [32] may improve model prediction of severe CDI in Black and other non-White populations.

Our results were generally agnostic to specific algorithms and performed equally well when using both linear and nonlinear machine learning approaches. However, XGBoost was a notable exception, which showed poor performance and potentially reflects overfitting of the models. Furthermore, as the random forest and stacked ensemble methods do not produce coefficients, they are inherently less interpretable, and transforming model results into a risk score is more complex compared with lasso regression [33].

In general, the variables that carried the greatest importance were consistent across models. Importantly, all variables were collected within 48 hours of CDI diagnosis, which increases the clinical utility of our model by allowing for early risk stratification of patients. White blood cell count >15 000 cells/mL and acute rise in serum creatinine >1.5 mg/dL are well-established markers for severe CDI [34]. In addition, older age [35, 36], hypoalbuminemia [37], low hemoglobin levels [38], concurrent antibiotic use [37], and ICU admission [13, 37] have also previously been identified as predictors of poor outcomes from CDI. ICU admission is likely a marker of more severe disease, as higher rates of comorbid health conditions, laboratory derangements, and antibiotic use were observed in this group. As our model identified similar predictors of importance as other studies, we suspect that employing a large, multicenter cohort and using best practices for predictive modeling may have allowed for better model parameterization and optimal variable selection, ultimately resulting in improved model performance compared with prior scoring systems [14].

Our study has several notable strengths. To our knowledge, this is the largest study to combine data from 4 distinct cohorts composed of both temporally and geographically distinct patients. By employing and comparing several well-validated predictive modeling techniques, we were able to develop a highly accurate model for predicting complicated CDI using readily available structured EHR data. In addition, the use of permutation-based variable importance analysis allowed us to identify the importance of each predictor in our models.

However, there are several limitations as well. First, there were significant differences in model performance in White vs non-White populations, which suggests that inadvertent biases were encoded in the machine learning algorithms. Second, as our analysis was retrospective, some model features may not be true risk factors but rather markers for the beginning of complicated CDI itself. Third, our models were derived from 3 academic medical centers in the Midwestern United States. Model performance will have to be evaluated outside of this setting to confirm generalizability. Fourth, we cannot exclude the possibility that patients may have experienced the outcome at another hospital where we do not have records, and thus we may have potentially underestimated the extent of complications from CDI. In addition, although CDI testing was recommended only for symptomatic patients and this was further validated by chart review, some positive CDI tests might still reflect asymptomatic carriers. However, PCR-based testing strategies were most likely to pick up asymptomatic carriage, while our sensitivity analyses showed that model performance remained robust when stratified by testing strategy. Fifth, as attributable CDI was adjudicated by each site investigator, there may have been site-specific differences in attributable CDI. However, definitions for attributable CDI were specifically defined a priori, while our sensitivity analyses suggest that models were robust despite any potential differences across sites. Lastly, our model employed a large number of variables, which may affect clinician-perceived usability. However, as these variables are all easily accessible within the EHR, we anticipate that an automated decision tool, embedded directly within the EHR, would minimize these concerns.

In summary, in a large heterogeneous population from a multicenter cohort, we demonstrated that machine learning algorithms based on structured EHR data can accurately estimate patients' risk for disease-related complications from CDI. Our approach leverages variables that can be readily extracted from the EHR once a diagnosis of CDI has been made. Future studies may determine whether prospective deployment of this model may aid clinicians to tailor patient therapy in real time and allow for early use of more aggressive therapies to minimize risk of complications. However, our results and sensitivity analysis demonstrated implicit racial biases. While model bias and unfairness have the potential to

exacerbate racial inequities in health care, they can also be used to overcome inequalities by proactively mitigating existing disparities [25, 26]. By identifying clear racial differences in model performance, we can apply group-specific modifications of decision thresholds to ensure fairness. The goals of future studies should be to reduce the disparity in model accuracy between White and non-White patients and to improve performance overall, and both goals could possibly be served by including host- or microbe-derived biomarkers alongside clinical data [32].

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

Financial support. This study was supported by National Institutes of Health grants HS027431 (to K.R.) and DK124567 (to A.A.L.).

Potential conflicts of interest. P.D.R.H. received consulting fees from AbbVie, Amgen, Genentech, JBR Pharma, and Lycera. J.A.B. received consulting fees from Buhlmann Diagnostic Corp. K.R. is supported in part by an investigator-initiated grant from Merck & Co, Inc.; he has consulted for Bio-K+ International, Inc., Roche Molecular Systems, Inc., Seres Therapeutics, Inc., and Summit Therapeutics, Inc. All other authors report no disclosures.

Author contributions. Study concept and design: A.L., K.R., P.D.R.H. Acquisition: D.A.P., D.M., D.S., P.P. Analysis or interpretation of data: all authors. Drafting of the manuscript: all authors. Figures: J.A.B., C.A.S., S.R., A.L., K.R. Critical revision of the manuscript: all authors. Final approval: all authors.

Data availability. All data produced in the present study are available upon reasonable request to the authors.

References

1. Hall AJ, Curns AT, McDonald LC, Parashar UD, Lopman BA. The roles of *Clostridium difficile* and norovirus among gastroenteritis-associated deaths in the United States, 1999–2007. Clin Infect Dis **2012;** 55:216–23.
2. Lessa FC, Mu Y, Bamberg WM, et al. Burden of *Clostridium difficile* infection in the United States. N Engl J Med **2015;** 372:825–34.
3. Cheng YW, Fischer M. Treatment of severe and fulminant *Clostridioides difficile* infection. Curr Treat Options Gastroenterol **2019;** 17:524–33.
4. Dubberke ER, Olsen MA. Burden of *Clostridium difficile* on the healthcare system. Clin Infect Dis **2012;** 55(Suppl 2):S88–92.
5. Belmares J, Gerding DN, Parada JP, Miskevics S, Weaver F, Johnson S. Outcome of metronidazole therapy for *Clostridium difficile* disease and correlation with a scoring system. J Infect **2007;** 55:495–501.
6. Bhangu S, Bhangu A, Nightingale P, Michael A. Mortality and risk stratification in patients with *Clostridium difficile*-associated diarrhoea. Colorectal Dis **2010;** 12:241–6.
7. Bloomfield MG, Carmichael AJ, Gkrania-Klotsas E. Mortality in *Clostridium difficile* infection: a prospective analysis of risk predictors. Eur J Gastroenterol Hepatol **2013;** 25:700–5.
8. Butt E, Foster JA, Keedwell E, et al. Derivation and validation of a simple, accurate and robust prediction rule for risk of mortality in patients with *Clostridium difficile* infection. BMC Infect Dis **2013;** 13:316.
9. Drew RJ, Boyle B. RUWA scoring system: a novel predictive tool for the identification of patients at high risk for complications from *Clostridium difficile* infection. J Hosp Infect **2009;** 71:93–4. Author reply 4–5.
10. Gujja D, Friedenberg FK. Predictors of serious complications due to *Clostridium difficile* infection. Aliment Pharmacol Ther **2009;** 29:635–42.
11. Hensgens MPM, Dekkers OM, Goorhuis A, LeCessie S, Kuijper EJ. Predicting a complicated course of *Clostridium difficile* infection at the bedside. Clin Microbiol Infect **2014;** 20:O301–8.
12. Na X, Martin AJ, Sethi S, et al. A multi-center prospective derivation and validation of a clinical prediction tool for severe *Clostridium difficile* infection. PLoS One **2015;** 10:e0123405.
13. Toro DH, Amaral-Mojica KM, Rocha-Rodriguez R, Gutierrez-Nuñez J. An innovative severity score index for *Clostridium difficile* infection: a prospective study. Infect Dis Clin Pract **2011;** 19:336–9.
14. Perry DA, Shirley D, Micic D, et al. External validation and comparison of *Clostridioides difficile* severity scoring systems. Clin Infect Dis **2022**; 74:2028–35.
15. Rao K, Micic D, Natarajan M, et al. *Clostridium difficile* ribotype 027: relationship to age, detectability of toxins A or B in stool with rapid testing, severe infection, and mortality. Clin Infect Dis **2015;** 61:233–41.
16. Stekhoven DJ, Bühlmann P. Missforest–non-parametric missing value imputation for mixed-type data. Bioinformatics **2012;** 28:112–8.
17. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open **2013;** 3:e002847.
18. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Available at: https://www.tidymodels.org Accessed 21 June 2022.
19. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw **2010;** 33:1–22.
20. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw **2017;** 77:1–17.
21. 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 *for Computing Machinery*. **2016**:785–94. doi:10.1145/2939672.2939785.
22. Couch S, Kuhn M. Stacks: tidy model stacking. Available at: https://stacks.tidymodels.org/ Accessed 21 June 2022.
23. Greenwell BM, Boehmke BC. Variable importance plots—an introduction to the vip package. R J **2020;** 12:343–66.
24. Staniak M, Biecek P. Explanations of model predictions with live and breakDown packages. R J **2019;** 10:395–409.
25. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health **2022;** 4:e406–14.
26. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med **2021;** 27:136–40.
27. Madden GR, Petri WA, Costa DVS, Warren CA, Ma JZ, Sifri CD. Validation of clinical risk models for *Clostridioides difficile* attributable outcomes. Antimicrob Agents Chemother 2022**;** 66:e00676–22.
28. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron **1976;** 16:31–41.
29. Jones CA, McQuillan GM, Kusek JW, et al. Serum creatinine levels in the US population: Third National Health and Nutrition Examination Survey. Am J Kidney Dis **1998;** 32:992–9.
30. Hsu CY, Yang W, Parikh RV, et al. Race, genetic ancestry, and estimating kidney function in CKD. N Engl J Med **2021;** 385:1750–60.
31. Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med **2021;** 385:1737–49.
32. Ressler A, Wang J, Rao K. Defining the black box: a narrative review of factors associated with adverse outcomes from severe *Clostridioides difficile* infection. Therap Adv Gastroenterol **2021;** 14:175628482110481.
33. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Github; **2022**. Available at: christophm.github.io/interpretable-ml-book/ Accessed 21 June 2022.
34. 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 **2018;** 66:987–94.
35. Ananthakrishnan AN, Guzman-Perez R, Gainer V, et al. Predictors of severe outcomes associated with *Clostridium difficile* infection in patients with inflammatory bowel disease. Aliment Pharmacol Ther **2012;** 35:789–95.
36. Chiang HY, Huang HC, Chung CW, et al. Risk prediction for 30-day mortality among patients with *Clostridium difficile* infections: a retrospective cohort study. Antimicrob Resist Infect Control **2019;** 8:175.
37. Tay HL, Chow A, Ng TM, Lye DC. Risk factors and treatment outcomes of severe *Clostridioides difficile* infection in Singapore. Sci Rep **2019;** 9:13440.
38. Lee E, Song KH, Bae JY, et al. Risk factors for poor outcome in community-onset *Clostridium difficile* infection. Antimicrob Resist Infect Control **2018;** 7:75.