BMJ Open Gastroenterology

Prediction of in-hospital mortality of *Clostriodiodes difficile* infection using critical care database: a big data-driven, machine learning approach

Hao Du ,[1] Kewin Tien Ho Siah ,[2,3] Valencia Zhang Ru-Yan,[3] Readon Teh ,[3] Christopher Yu En Tan,[2] Wesley Yeung,[3,4] Christina Scaduto,[5] Sarah Bolongaita,[5] Maria Teresa Kasunuran Cruz,[3] Mengru Liu,[6] Xiaohao Lin,[7] Yan Yuan Tan,[8] Mengling Feng[1,9]

To cite: Du H, Siah KTH, Ru-Yan VZ, *et al*. Prediction of in-hospital mortality of *Clostriodiodes difficile* infection using critical care database: a big data-driven, machine learning approach. *BMJ Open Gastro* 2021;**8**:e000761. doi:10.1136/bmjgast-2021-000761

HD and KTHS contributed equally.

Received 5 August 2021
Accepted 5 October 2021

For numbered affiliations see end of article.

Correspondence to
Dr Kewin Tien Ho Siah; kewin_siah@nuhs.edu.sg

ABSTRACT

Research objectives *Clostriodiodes difficile* infection (CDI) is a major cause of healthcare-associated diarrhoea with high mortality. There is a lack of validated predictors for severe outcomes in CDI. The aim of this study is to derive and validate a clinical prediction tool for CDI in-hospital mortality using a large critical care database.

Methodology The demographics, clinical parameters, laboratory results and mortality of CDI were extracted from the Medical Information Mart for Intensive Care-III (MIMIC-III) database. We subsequently trained three machine learning models: logistic regression (LR), random forest (RF) and gradient boosting machine (GBM) to predict in-hospital mortality. The individual performances of the models were compared against current severity scores (Clostridiodes difficile Associated Risk of Death Score (CARDS) and ATLAS (Age, Treatment with systemic antibiotics, leukocyte count, Albumin and Serum creatinine as a measure of renal function) by calculating area under receiver operating curve (AUROC). We identified factors associated with higher mortality risk in each model.

Summary of results From 61 532 intensive care unit stays in the MIMIC-III database, there were 1315 CDI cases. The mortality rate for CDI in the study cohort was 18.33%. AUROC was 0.69 (95% CI, 0.60 to 0.76) for LR, 0.71 (95% CI, 0.62 to 0.77) for RF and 0.72 (95% CI, 0.64 to 0.78) for GBM, while previously AUROC was 0.57 (95% CI, 0.51 to 0.65) for CARDS and 0.63 (95% CI, 0.54 to 0.70) for ATLAS. Albumin, lactate and bicarbonate were significant mortality factors for all the models. Free calcium, potassium, white blood cell, urea, platelet and mean blood pressure were present in at least two of the three models.

Conclusion Our machine learning derived CDI in-hospital mortality prediction model identified pertinent factors that can assist critical care clinicians in identifying patients at high risk of dying from CDI.

Summary box

What is already known about this subject?

- *Clostriodiodes difficile* infection (CDI) is one of the most common hospital-acquired infections with high mortality rates.
- Several attempts have been made to develop models to predict CDI severity or mortality. However, they are less than ideal due to a lack of routinely recordable variables, low level of discrimination and limited subgroup applicability.

What are the new findings?

- Machine learning models are developed to predict in-hospital mortality of patients with CDI.
- The proposed machine learning models outperformed existing severity scores in predicting mortality outcomes.

How might it impact on clinical practice in the foreseeable future?

- The proposed machine learning models could incorporate variability in laboratory data and comorbidities into prediction.
- The proposed models can facilitate early recognition of CDI severity and enable timely intervention to patients in need.

INTRODUCTION

Clostridiodes difficile infection (CDI) has been recognised as a major cause of healthcare-associated diarrhoea in adult patients.[1] As one of the rising healthcare-associated infections worldwide, it causes a significant burden on hospital resources. The reason for the rise in CDI is largely due to the increasing use of antibiotics in current clinical practice, as well as an ageing patient population in the hospitals.[2] Consequently, the disease burden of CDI has been rising, with more elderly patients facing longer hospitalisations, higher healthcare costs, as well as more severe morbidity and mortality.[3]

C. difficile is transmitted by contact with infected faecal material or spores which can survive in the environment for several

months.[4] *C. difficile* is easily passed on via the hands of healthcare workers.[5] Antibiotic use is a major risk factor for CDI, causing alterations in gut microbiota that protect against gut infection, resulting in proliferation of *C. difficile*.[6] The primary mediators of inflammation in CDI are large clostridial toxins, toxin A (TcdA) and toxin B (TcdB), which bind to and enter the colonic epithelium. This results in a sequence of host cellular responses to cause diarrhoea, inflammation and tissue necrosis.[7] This manifests clinically as asymptomatic colonisation, mild diarrheal illness or more severe disease, including pseudomembranous colitis, toxic megacolon, sepsis and in severe cases, death.[8]

Treatment recommendations for CDI vary according to disease severity, ranging from oral antibiotics to surgical intervention.[9] Metronidazole and vancomycin remain the cornerstone of CDI treatment, while fidaxomicin, a newly approved drug, is a new alternative. In patients with severe CDI, early surgical consultation is recommended by the World Society of Emergency Surgery and the Infectious Diseases Society of America—Society for Healthcare Epidemiology of America.[10] Prompt subtotal or total colectomy can reduce mortality[11] in patients with megacolon, colonic perforation or for patients with septic shock and associated organ failure.[12]

Despite the increasing prevalence of CDI in the developed world, validated methods to predict severe disease have not been established.[13] We recently published a systematic review of severe CDI predictors, but found that present risk scoring systems have been limited by small sample size and heterogeneity in definition of severe CDI.[14] Proposed severity scores such as Clostridiodes difficile Associated Risk of Death Score (CARDS)[15] and ATLAS Score[16 17] (combination of age, treatment with systemic antibiotics, leucocyte count, serum albumin and serum creatinine) have not been widely adopted in current clinical practice.

The aim of this study is to derive and validate a clinical prediction tool for severe outcomes in CDI. We standardised our measured outcome in this study as mortality to create a straightforward model that predicts for in-hospital CDI mortality. We sought to address limitations of existing severity scores by developing our risk prediction model from a large database, the Medical Information Mart for Intensive Care-III (MIMIC-III)—an open-source, reputable and repeatable electronic-intensive care unit database.

METHODS

Data source and extraction

This was a retrospective study, and all patients were de-identified. Thus, informed consent was waived by the ethics committee of Beth Israel Deaconess Medical Center. Data were extracted from MIMIC-III using structure query language (SQL) with PostgreSQL 11.5 (PostgreSQL Global Development Group).[18] The MIMIC-III database contains health-related data associated with over 40 000 patients between 2001 and 2012, and it is publicly available.

It comprises health data of over 40 000 patients who stayed in intensive care units (ICUs) of the Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA between 2001 and 2012. The database is comprehensive in nature and includes patient vital sign measurements at 1-hour intervals, demographics, laboratory test results, procedures and caregiver notes. Vincent *et al*[19] and Gehrmann *et al*[20] are notable studies that have leveraged on the MIMIC-III database for large-scale retrospective analyses. Our data-driven approach allows us to develop a CDI severity prediction tool based on clinical outcomes rather than existing literature, eliminating the risk of bias and welcoming new possibilities of CDI severity predictors. Additionally, patients from MIMIC-III are managed in the ICU and thus have comprehensive clinical and biochemical data available, allowing for novel variables to be taken into account when searching for CDI severity predictors.

Inclusion criteria and definition

The patients were extracted based on The International Classification of Diseases, ninth Revision, Clinical Modification code of '008.45', indicating the diagnosis of 'intestinal infection due to *Clostridium difficile*' at hospital discharge. We also extracted data for the first ICU stay of patients aged between 16 and 90 years old. Other extracted information were patients' demographics (age, gender, comorbidities of diabetes mellitus, chronic kidney disease and chronic ischaemic heart disease), laboratory test results (anion gap, albumin, bicarbonate, bilirubin, creatinine, chloride, glucose, haematocrit, haemoglobin, lactate, platelet, potassium, partial thromboplastin time, international normalised ratio, prothrombin time, sodium, blood urea nitrogen (BUN), white blood cell (WBC) count, calcium, free calcium) and vital signs (heart rate, respiratory rate, SpO_2, body temperature, systolic blood pressure, diastolic blood pressure, mean arterial pressure). For patients with multiple measurements of laboratory tests and vital signs, we only kept the results of the first measurement, where the measurement time is closest to the patients' ICU admission time. Patients' missing information for laboratory tests and vital signs were imputed with the mean value for each variable in the cohort. Imputation of missing values is crucial for modelling. Removing records with missing values and some other filtering methods have been shown to introduce bias, which affects the performance of models in many ways, thereby limiting their generalisation.[21] Imputation with mean values is commonly adopted as it maintains the distribution of predictors.[22–24] We also implemented the RidgeRegression model for data imputation with reference to Cosgriff *et al*[25] and similar results were obtained. The endpoint of this study is in-hospital mortality.

Statistical analysis

We trained three machine learning (ML) models: logistic regression (LR),[26] random forest (RF)[27] and

gradient boosting machine (GBM)[28] to predict in-hospital mortality based on the clinical features that were commonly measured on patients' ICU admissions.

LR is commonly used in clinical research to model disease presence (diagnosis) or disease outcomes (prognosis). In our study, we used LR to predict the chance of the outcome based on the characteristics of the risk factors (predictors). A linear combination of predictors is used to fit a 'logit' transformation of the probability of the outcome. In the LR model, we reported the results using OR and the corresponding 95% CIs for all covariates.

RF uses random bootstrap samples of raw data samples to construct a series of decision trees and use them for medical prediction and classification tasks. It is a non-parametric classifier that constructs a hierarchical decision tree by splitting the data between the categories of outcome variables at a given step (node) according to the 'if-then' rule of a given set of risk factors. The model repeats it into two subnodes, which come from the root node that contains the entire sample (for a detailed description of RF, please refer to Breiman[27]). This 'ensemble learning' classification method could reduce prediction variance and prevent overfitting to training sets.

Similar to RF, GBM model is also a decision tree based approach. Boosting is a group of methods which combine weak learners into strong learners. In decision tree boosting, each decision tree is trained on a subset of original dataset. For example, the first decision tree assigns equal weights to each observation and fits on the equally weighed dataset. After the first decision tree is evaluated, the boosting model re-weights each observation. The weights of difficult cases are increased and the weights of easy cases are decreased. The following decision tree is fitted on this re-weighted data. In this way, the performance of the overall model is improved based on the predictions of the first decision tree. The boosting model is now an ensemble of the first and second decision tree. Next, we evaluate the classification error of the boosting ensemble model and fit the third decision tree to predict the revised residuals. The process is repeated for a defined number of iterations. The new decision trees improve the ensemble model by fitting on the observations that are incorrectly predicted by previous decision trees. The final ensemble model is predicted by the weighted sum of the predicted values of all fitting decision trees. In GBM, particularly, the model uses a loss function to identify weak learners and gradients to minimise the loss.

All three models used the same set of training and testing data. We split the original dataset into 80% train-set and 20% test-set, in which we ensure the proportion of the positive outcomes was the same in both sets by stratifying the dataset based on hospital mortality. An RF-based feature selection method[29] was used to detect key features for mortality prediction. The selected features were adjusted as covariates in the three models.

Model performance metrics

In our study, all the models predict the probability of in-hospital mortality for each patient's ICU stay. We then use this probability as a risk score for clinicians to better understand the overall risk of death for individual patients at admission. If the risk score exceeds a specific threshold, the patient would be classified to a high risk group and receive attention from clinicians in advance.

We plotted receiver operating curve (ROC) according to different selected thresholds and calculated the mean area under ROC (AUROC) to evaluate the performance for each model. To ensure the robustness of our finding, we calculated the 95% CIs of the AUROC with 100 bootstraps of the train-test split. The way of calculating the bootstrapped CIs is inspired by Oh *et al*,[30] where the authors selected the 95th percentile of the predicted probability as the decision threshold for prediction of CDI diagnosis. We computed the AUROC of other proposed severity scores, CARDS and ATLAS, to compare with our model performance. We also computed and compared the selected threshold with accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for all the three models.

In addition, calibration is an important measure of predictive models. Calibration measures the model's ability to produce predictions that are averagely close to the average observed result. For example, a model is said to be well-calibrated, if for 100 patients with a predicted risk of x%, close to x patients have developed the outcome. We used fivefold cross validation to calibrate each model: for each fold, we used the train set to fit the model and calibrate the trained model on the test set. The probabilities for each of the folds are then averaged for prediction.

Besides, we further investigated features' importance in each model. For the LR model, we identified statistically significant features with p value <0.05. We ranked the features by the absolute value of their coefficients and obtained the top significant features for LR. For both RF and GBM models, feature importance was determined by counting the times (normalised) when the feature was chosen by the model to split the nodes in the decision trees. The top features are those with more counts. The statistical analyses were performed in Python 3.6. The codes are publicly available.[31]

RESULTS

From 61 532 ICU stays in the MIMIC-III database, we identified 1315 unique ICU stays with diagnosis of CDI. Demographics and clinical characteristics of the study population are provided (table 1). The mortality rate in the study cohort was 18.33% (241 over 1315 ICU stays). For each bootstrap iteration, train and test set included 1052 and 263 unique ICU stays, respectively. On the test set, CARDS achieved AUROC of 0.57 (95% CI, 0.51 to 0.65) and ATLAS Score achieved 0.63 (95% CI, 0.54 to 0.70). For our proposed models, the LR model achieved

Table 1 The basic characteristics of the study cohort

Patient demographics and clinical features	Number of missing values (proportion (%))	Median (IQR) or number of non-null values (proportion (%))
Age	0 (0)	70.00 (58.14–79.48)
Gender (male)	0 (0)	649 (49.35)
Comorbidities (diagnosed with diabetes mellitus, chronic ischaemic heart disease or chronic kidney disease)	0 (0)	482 (36.65)
Anion gap	2 (0.15)	15.00 (13.00–18.00)
Albumin	305 (23.2)	2.70 (2.30–3.20)
Bicarbonate	2 (0.15)	23.00 (20.00–27.00)
Bilirubin	256 (19.47)	0.50 (0.30–0.95)
Creatinine	2 (0.15)	1.30 (0.80–2.40)
Chloride	2 (0.15)	103.00 (99.00–107.00)
Glucose	2 (0.15)	128.00 (101.00–165.00)
Haematocrit	2 (0.15)	32.00 (28.40–36.00)
Haemoglobin	2 (0.15)	10.50 (9.30–11.90)
Lactate	217 (16.50)	1.80 (1.30–2.80)
Platelet	2 (0.15)	243.00 (164.00–365.00)
Potassium	2 (0.15)	4.20 (3.70–4.70)
Partial thromboplastin time (PTT)	39 (2.97)	31.40 (27.00–38.20)
International normalised ratio (INR)	35 (2.66)	1.30 (1.20–1.70)
Prothrombin time (PT)	35 (2.66)	14.70 (13.40–17.90)
Sodium	2 (0.15)	138.00 (135.00–141.00)
Blood urea nitrogen (BUN)	2 (0.15)	27.00 (17.00–46.00)
White blood cells (WBCs)	2 (0.15)	12.90 (8.60–19.60)
Calcium	35 (2.66)	8.30 (7.70–8.90)
Free calcium	734 (55.82)	1.10 (1.02–1.17)
Heart rate	13 (0.99)	95.00 (81.00–110.00)
Respiratory rate	13 (0.99)	20.00 (16.00–24.00)
Oxygen saturation (SpO_2)	14 (1.06)	98.00 (95.00–100.00)
Temperature (°C)	16 (1.22)	36.67 (36.06–37.33)
Systolic blood pressure	13 (0.99)	118.00 (103.00–138.00)
Diastolic blood pressure	13 (0.99)	60.00 (51.00–72.00)
Mean arterial pressure	13 (0.99)	77.00 (66.08–89.58)

a mean AUROC of 0.69 (95% CI, 0.60 to 0.76). RF model achieved AUROC of 0.710 (95% CI, 0.620 to 0.770) and GBM model achieved AUROC of 0.720 (95% CI, 0.64 to 0.78) (figure 1). The calibration was evaluated for each model. In one bootstrapping test set, for example, all of three models demonstrated good calibrations. The Brier scores for LR, RF and GBM models were 0.139, 0.131 and 0.132, respectively.

The decision threshold of the high risk group was selected based on the 95th percentile of predicted risk scores. LR achieved accuracy of 79.85%, sensitivity of 10.20%, specificity of 95.79% and PPV of 35.71%. RF model achieved accuracy of 84.41%, sensitivity of 22.45%, specificity of 98.60% and PPV of 78.57%. GBM model obtained accuracy of 82.13%, sensitivity of 16.33%, specificity of 97.20% and PPV of 57.14% (figure 2).

We observed similarities across the models in terms of the top 10 significant features (table 2). Albumin, lactate and bicarbonate were significant for all the models. Free calcium, potassium, WBC, BUN, platelet and mean blood pressure were also important features, which were agreed by two of the three models. Gender and some lab tests results (haemoglobin, hematocrit, anion gap, creatinine) were considered important in the LR model but not in the other two models. Other variables, including heart rate, age, respiratory rate and sodium appeared only once in the list of either RF or GBM but not in LR model.

DISCUSSION

In this cohort study, we sought to employ ML in developing a big data-based prediction model to predict in-hospital mortality of patients with CDI admitted to the ICU. All three of our advanced ML algorithms accurately predict the probability of in-hospital mortality for each patient's ICU stay. All ML models had adequate discrimination (ie, AUROC between 0.69 and 0.72) in predicting patient mortality. Our AUROC was comparable to that of the

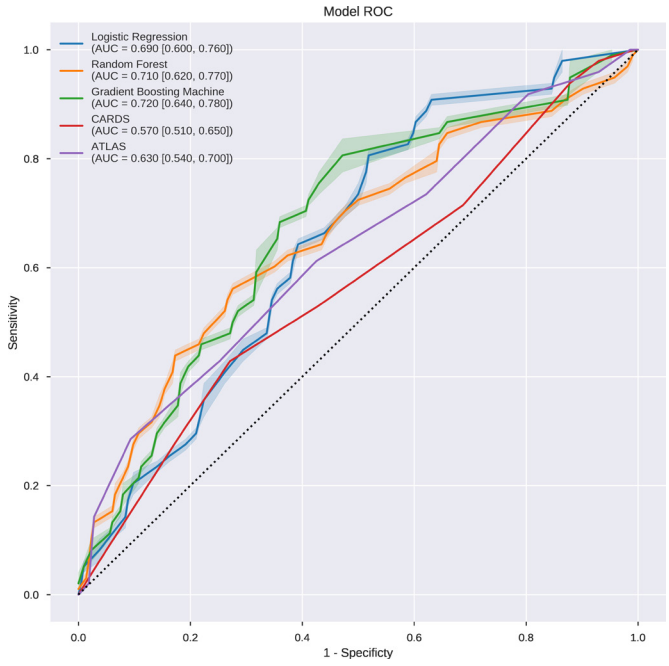


Figure 1 Discriminative performance of the models on the test set. The receiver operating characteristics curves illustrate the trade-off in performance between the false-positive rate (1–specificity) and the true-positive rate (sensitivity). Three models achieved good discriminative performance as measured by the area under the ROC curve (AUROC): logistic regression at 0.69, random forest at 0.71, GBM at 0.72. AUC, Area Under the Curve; ATLAS, Age, Treatment with systemic antibiotics, Leucocyte count, Albumin and Serum creatinine as a measure of renal function; CARDS, Clostridiodes difficile Associated Risk of Death Score; GBM, gradient boosting machine; ROC, receiver operating curve.

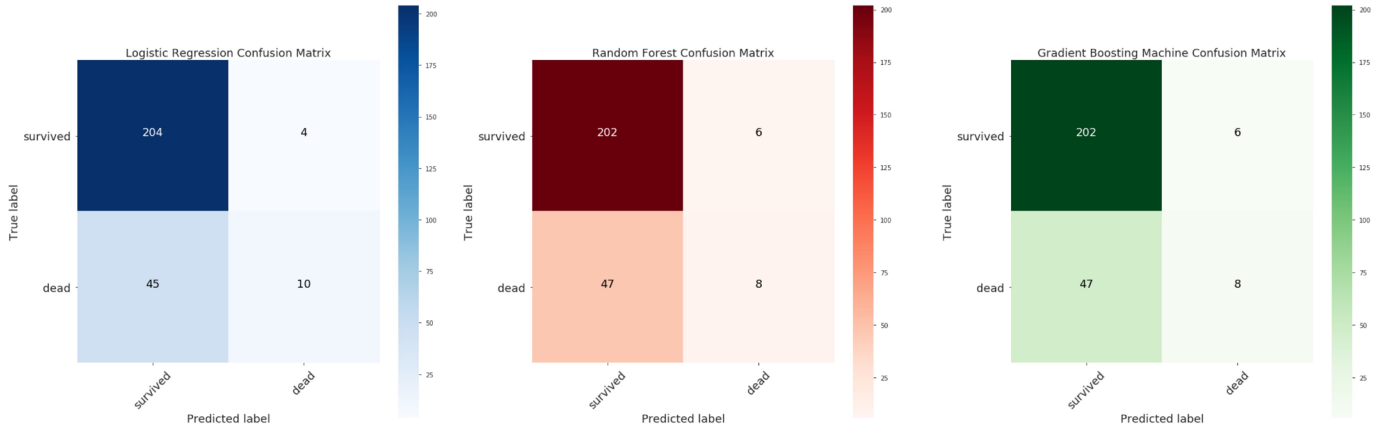
Table 2 The top 10 risk/protective factors across three models, ranked from most important to least important

Logistic regression	Random forest	GBM
Free calcium	White blood cell	White blood cell
Gender	Blood urea nitrogen	Bicarbonate
Haemoglobin	Platelet	Mean blood pressure
Albumin	Albumin	Blood urea nitrogen
Potassium	Mean blood pressure	Albumin
Haematocrit	Lactate	Lactate
Lactate	Bicarbonate	Platelet
Bicarbonate	Heart rate	Respiratory rate
Anion gap	Age	Sodium
Creatinine	Free calcium	Potassium

GBM, gradient boosting machine.

CARDS proposed by Kassam *et al*[15], which was 0.77. Though the PPV was low for our ML models, the specificity of the three models was high, ranging from 95.79% to 98.60%.

By using RF model and 95% percentile threshold, the cohort can be stratified into high-risk and low-risk groups. In the high-risk group, 100% of the patients had albumin values beyond normal range (3.4 to 5.4 g/dL); 78.57% of the patients had lactate values beyond normal range (0.5 to 2.2 mmol/L); 100% of the patients had bicarbonate values beyond normal range (23 to 30 mEq/L). In the low-risk group, 78.31% of the patients had albumin value beyond normal range; 52.61% of the patients had lactate values out of normal range; 64.66% of the patients had bicarbonate



Threshold: 0.34
Accuracy: 81.37%
Sensitivity: 18.18%
Specificity: 98.08%
PPV: 71.43%
NPV: 81.93%

Threshold: 0.40
Accuracy: 79.85%
Sensitivity: 14.55%
Specificity: 97.12%
PPV: 57.14%
NPV: 81.12%

Threshold: 0.40
Accuracy: 82.58%
Sensitivity: 18.18%
Specificity: 97.84%
PPV: 66.67%
NPV: 83.46%

Figure 2 Confusion matrices of logistic regression (left), random forest (middle) and GBM (right) on test set. Selecting a decision threshold based on the 95th percentile results in classifiers that achieved good specificity of above 95%. GBM, gradient boosting machine; NPV, negative predictive value; PPV, positive predictive value.

values out of normal range. The mortality of high-risk group is 57.14% while the mortality rate of low-risk group is 18.87%.

We also found several variables which were not typically incorporated into risk scores such as platelet, in addition to more established predictors such as albumin level, BUN level and WBC count. Our ML models could incorporate variability in laboratory data and many comorbidities into prediction, which other standard prognostic tools are unable to perform.

One advantage of ML classifiers, such as random-forest approaches, over purely regression-based classifiers, is that ML can take into account unexpected predictor variables and possible connections.[32] There can be many potential predictors, especially with increasing use of electronic health records, which may be overlooked with a predefined hypothesis.[33 34] An advanced ML approach allows for evaluation of far more clinical variables than would be present in traditional modelling approaches. Hence ML algorithms can promote identification of clinically important variables in patients with *C. difficile* which may not be recognised with a more conventional approach.

In our study, we found that albumin, lactate and bicarbonate were significant across all models. The finding of serum albumin being a predictor of mortality is concordant with our previous systematic review, which showed that at least five of the 31 articles identified prior comorbidities, age, white blood cell count, serum albumin, serum creatinine and ICU admission as predictors of severity.[14] Interestingly, none of the 31 studies on *C. difficile* severity predictors included lactate and bicarbonate levels. Furthermore, all international clinical guidelines on severe CDI, such as American College of Gastroenterology and European Society of Clinical Microbiology and Infectious Diseases, do not include lactate and bicarbonate levels. Elevated lactate levels are known to be significantly associated with in-hospital mortality and are featured in the surviving sepsis bundle.[35] Likewise, low bicarbonate and anion gap, which is observed in metabolic acidosis, has been used in assessing in-hospital mortality for patients admitted to the ICU[36] for patients with acute pancreatitis[37] or cardiogenic shock.[38] These biomarkers, when taken together and weighted according to their importance in our prediction models, can tell us more about the condition of a patient than just one biomarker alone. Similarly, factors which predicted mortality in CDI in two out of three of our models, namely free calcium, potassium and lactate, were not mentioned by any of the 31 studies. This could be due to the lack of availability of such data in the wards.

Our study is unique as it assesses the patient's parameters at the point of admission to the ICU instead of the point of diagnosis of *C. difficile*, as proposed by other studies. Analysing patients' data at ICU admission would be representative of the patient population whose management can best benefit from our study, as data analysed are reflective of patients with severe CDI requiring ICU management. The ultimate aim of our proposed ML model is to prognosticate patients with CDI and to catch those whose conditions are likely to worsen early on.

We recognise the limitations faced by our study. Data were retrospectively extracted from the MIMIC-III database, an electronic health record of a single academic medical centre in the USA, which may result in concerns regarding the generalisability of conclusions. We attempted to alleviate this limitation by evaluating our models with 100 bootstrapping iterations and obtaining CIs for each model. The data and codes are publicly available to researchers to replicate the study and evaluate the generality of the proposed models. Inherent to the retrospective nature of the study, we face selection bias as the majority of the population are Caucasian with few African Americans and Asians. The future plan of our study is to conduct prospective research to understand the real-time performance of proposed models. In addition, as data were collected over the duration of 2001–2012, treatment and practices may vary from the current standard. However, as the pathophysiology of progression in CDI is likely to remain unchanged, the clinical progression and laboratory values of these patients remain applicable. Another limitation is that the decision threshold of prediction was selected based on 95th percentile of the predicted probability.[30] In future studies, cross-validation methods can be used to select the optimal percentile and decision threshold in a data-driven manner. Imputation of missing values can be another limitation of this study. The imputation of mean values may not provide utility in the clinical settings. Regression-based imputation method such as ridge regression could be used as an alternative method. In our experiments, we obtained similar results when the missing values were imputed by ridge regression model. Furthermore, external validation of the model was not performed. External validation in a separate, independent dataset is considered important in fully evaluating the performance of prognostic models, and will be the direction of our future research.

In conclusion, by learning from the shortcomings of previous severity models, we have employed a robust and objective ML approach, while capitalising on one of the most extensive ICU databases to develop a CDI severity prediction model. This can potentially transform hospital care of patients by alerting clinicians of deteriorations and making timely intervention available to patients. Further exploration in clinical studies would be necessary to verify and refine our CDI predictor.

Author affiliations
[1]Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore
[2]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3]University Medicine Cluster, National University Hospital, Singapore
[4]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[5]Department of Global Health and Population, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA
[6]School of Computing and Information Systems, Singapore Management University, Singapore
[7]Machine Intellection Department, Institute for Infocomm Research, Agency for Science Technology and Research, Singapore
[8]Alliance Healthcare Group, Singapore

[9]Institute of Data Science, National University of Singapore, Singapore

Contributors HD contributed to study design, data extraction, methods, data analysis and manuscript preparation. KTHS contributed to project director and principal investigator, study design, data interpretation and manuscript preparation. VZR-Y contributed to data interpretation and manuscript preparation. RT contributed to data interpretation and manuscript preparation. WY contributed to data extraction, data interpretation and data analysis. CS contributed to data extraction, data interpretation and data analysis. MTKC contributed to study design and data interpretation. ML contributed to data extraction and data analysis. XL contributed to data extraction and data analysis. YYT contributed to study design, data interpretation and data analysis. MF contributed to project director and principal investigator, study design, method and manuscript preparation. KTHS acting as guarantor. All authors have approved the final version of the manuscript submitted.

Funding This project is partially supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-100E-2020-055).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. MIMIC-III, a freely accessible critical care database. https://mimic.physionet.org/.

ORCID iDs
Hao Du http://orcid.org/0000-0002-5083-8122
Kewin Tien Ho Siah http://orcid.org/0000-0002-2795-5101
Readon Teh http://orcid.org/0000-0002-4230-063X

REFERENCES

1 Guery B, Galperine T, Barbut F. *Clostridioides difficile*: diagnosis and treatments. *BMJ* 2019;366:l4609.
2 Depestel DD, Aronoff DM. Epidemiology of Clostridium difficile infection. *J Pharm Pract* 2013;26:464–75.
3 Shorr AF, Zilberberg MD, Wang L, *et al*. Mortality and costs in Clostridium difficile infection among the elderly in the United States. *Infect Control Hosp Epidemiol* 2016;37:1331–6.
4 Hensgens MPM, Keessen EC, Squire MM, *et al*. Clostridium difficile infection in the community: a zoonotic disease? *Clin Microbiol Infect* 2012;18:635–45.
5 Johnson S, Gerding DN, Olson MM, *et al*. Prospective, controlled study of vinyl glove use to interrupt Clostridium difficile nosocomial transmission. *Am J Med* 1990;88:137–40.
6 Mullish BH, Williams HR. *Clostridium difficile* infection and antibiotic-associated diarrhoea. *Clin Med* 2018;18:237–41.
7 Smits WK, Lyras D, Lacy DB, *et al*. Clostridium difficile infection. *Nat Rev Dis Primers* 2016;2:16020.
8 Bartlett JG, Gerding DN. Clinical recognition and diagnosis of Clostridium difficile infection. *Clin Infect Dis* 2008;46 Suppl 1:S12–18.
9 Kelly CP LJ, Bakken JS. Clostridioides (formerly Clostridium) difficile infection in adults: treatment and prevention. UpToDate, 2019. Available: https://www.uptodate.com/contents/clostridioides-formerlyclostridium-difficile-infection-in-adults-treatment-and-prevention
10 Sartelli M, Di Bella S, McFarland LV, *et al*. 2019 update of the WSES guidelines for management of *Clostridioides* (*Clostridium*) *difficile* infection in surgical patients. *World J Emerg Surg* 2019;14:8.
11 Bagdasarian N, Rao K, Malani PN. Diagnosis and treatment of Clostridium difficile in adults: a systematic review. *JAMA* 2015;313:398–408.
12 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.
13 Beauregard-Paultre C, Abou Chakra CN, McGeer A, *et al*. External validation of clinical prediction rules for complications and mortality following Clostridioides difficile infection. *PLoS One* 2019;14:e0226672.
14 Zhang VRY, Woo ASJ, Scaduto C, *et al*. Systematic review on the definition and predictors of severe *Clostridiodes difficile* infection. *J Gastroenterol Hepatol* 2021;36:89–104.
15 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;43:725–33.
16 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.
17 Mulherin DW, Hutchison AM, Thomas GJ, *et al*. Concordance of the SHEA-IDSA severity classification for Clostridium difficile infection and the atlas bedside scoring system in hospitalized adult patients. *Infection* 2014;42:999–1005.
18 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:1–9.
19 Vincent J-L, Nielsen ND, Shapiro NI, *et al*. Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the MIMIC-III database. *Ann Intensive Care* 2018;8:107.
20 Gehrmann S, Dernoncourt F, Li Y, *et al*. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One* 2018;13:e0192360.
21 Winslow RL, Trayanova N, Geman D, *et al*. Computational medicine: translating models to clinical care. *Sci Transl Med* 2012;4:158rv1 1–rv11.
22 Critical Data M. *Secondary analysis of electronic health records*. Basingstoke: Springer Nature, 2016.
23 Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *Int J Med Inform* 2019;125:55–61.
24 Li F, Xin H, Zhang J, *et al*. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open* 2021;11:e044779.
25 Cosgriff CV, Celi LA, Ko S, *et al*. Developing well-calibrated illness severity scores for decision support in the critically ill. *NPJ Digit Med* 2019;2:1–8.
26 Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. Hoboken: John Wiley & Sons, 2013.
27 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
28 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001:1189–232.
29 Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw* 2010;36:1–13.
30 Oh J, Makar M, Fusco C, *et al*. A generalizable, data-driven approach to predict daily risk of Clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018;39:425–33.
31 Du H. Code for project 'Development of a risk prediction model for severe clostridiodes difficile infection', 2020. Available: https://github.com/DuHao10086/cdiff_mimic
32 Parikh RB, Manz C, Chivers C, *et al*. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw Open* 2019;2:e1915997.
33 Singal AG, Mukherjee A, Elmunzer BJ, *et al*. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol* 2013;108:1723–30.
34 Waljee AK, Higgins PDR, Singal AG. A primer on predictive models. *Clin Transl Gastroenterol* 2014;5:e44.
35 Casserly B, Phillips GS, Schorr C, *et al*. Lactate measurements in sepsis-induced tissue hypoperfusion: results from the surviving sepsis campaign database. *Crit Care Med* 2015;43:567–73.
36 Mehta HJ, Bhanusheli G, Nietert PJ, *et al*. Withdrawn: the association between initial anion gap and outcomes in medical intensive care unit patients. *J Crit Care* 2012. doi:10.1016/j.jcrc.2012.04.003. [Epub ahead of print: 12 Jun 2012].
37 Sharma V, Shanti Devi T, Sharma R, *et al*. Arterial pH, bicarbonate levels and base deficit at presentation as markers of predicting mortality in acute pancreatitis: a single-centre prospective study. *Gastroenterol Rep* 2014;2:226–31.
38 Wigger O, Bloechlinger S, Berger D, *et al*. Baseline serum bicarbonate levels independently predict short-term mortality in critically ill patients with ischaemic cardiogenic shock. *Eur Heart J Acute Cardiovasc Care* 2018;7:45–52.