Development of an Open-Access and Explainable Machine Learning Prediction System to Assess the Mortality and Recurrence Risk Factors of *Clostridioides Difficile* Infection Patients

Yui-Lun Ng, Michelle C. K. Lo, Kit-Hang Lee, Xiaochen Xie, Thomas N. Y. Kwong, Margaret Ip, Lin Zhang, Jun Yu, Joseph J. Y. Sung, William K. K. Wu, Sunny H. Wong, and Ka-Wai Kwok**

Identifying *Clostridioides difficile* infection (CDI) patients at risk of mortality or recurrence facilitates prevention, timely treatment, and improves clinical outcomes. The aim herein is to establish an open-access web-based prediction system, which estimates CDI patients' mortality and recurrence outcomes and explains machine learning prediction with patients' characteristics. Prognostic models are developed using four various types of machine learning algorithms and the statistical logistics regression model utilizing over 15 000 CDI patients from 41 hospitals in Hong Kong. The boosting-based machine learning algorithm gradient boosting machine (GBM) (Mortality AUC: 0.7878; Recurrence AUC: 0.7076) outperforms statistical models (Mortality AUC: 0.7573; Recurrence AUC: 0.6927) and other machine learning algorithms. As the difficulty to interpret complex machine learning results limits their use in the medical area, Shapley additive explanations (SHAP) are adapted to identify which features are crucial to the machine learning models and associate them with clinical findings. SHAP analysis shows that older age, reduced albumin levels, higher creatinine levels, and higher white blood cell count are the most highly associated mortality features, which is consistent with existing clinical findings. The open-access prediction system for clinicians to assess and interpret the risk factors of CDI patients is now available at https://www.cdiml.care/.

Predictive models that assess the mortality and recurrence risk of infectious diseases can assist clinicians in diagnosing and interpreting the critical risk factors for patient recovery. With the establishment of the electronic health records (EHR) systems, daily-acquired medical data, such as patients' medical history, laboratory test results, and medications, provide an abundant data source to further investigate the epidemiology of disease and develop more accurate, powerful, and robust prediction models. Statistical methods have served as the backbone for developing many clinical severity score indices or predictive models, as they are capable of identifying the key critical variables for clinical reference. However, statistical methods mainly focus on inference, which deduces the properties of the data related to the outcome effect, rather than optimizing the predictive power for patient outcomes.[1] Machine learning has proven its capability to develop highly precise prediction models in both medical and robotics applications such as EHR medical records systems,[2–4] robotics navigation systems,[5] and robotics sensing and control.[6–8] Apart from prediction accuracy, model interpretations are critical in clinical decision support but the difficulty in explaining complex machine learning model results limits their use in the medical area. To assist clinicians' diagnosis, revealing the key features leading to accurate machine learning models

Y. L. Ng, M. C. K. Lo, Dr. K. H. Lee, Dr. X. Xie, Dr. K. W. Kwok
Department of Mechanical Engineering
Faculty of Engineering
The University of Hong Kong
7/F, Haking Wong Building, Pokfulam Road, Hong Kong
E-mail: kwokkw@hku.hk

Dr. T. N. Y. Kwong, Dr. L. Zhang, Prof. J. Yu, Prof. J. J. Y. Sung,
Dr. W. K. K. Wu, Dr. S. H. Wong
SKL Laboratory of Digestive Disease
Li Ka Shing Institute of Health Sciences
Department of Medicine and Therapeutics
Faculty of Medicine
The Chinese University of Hong Kong
30-32 Ngan Shing Street, Shatin, NT, Hong Kong
E-mail: wonghei@cuhk.edu.hk

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202000188.

DOI: 10.1002/aisy.202000188

and associating them with existing clinical findings are one of our major objectives.

This study aims at estimating the mortality and recurrence outcomes of *Clostridioides difficile* infection (CDI) patients. CDI is the most common nosocomial enteric infection, and the symptoms of CDI patients can range from mild diarrhea to severe sepsis with organ failure, which may lead to significant morbidity and mortality.[9] Due to the high transmissibility of *C. difficile* and increased risk due to the widespread use of antibiotics, the disease carries a considerable health burden. We have previously reported that in Hong Kong, the incidence of CDI has increased by 26% from 15.41 cases to 36.31 cases per 100 000 persons from 2006 to 2014.[10] The approximated CDI incidences in the USA in 2011 was almost half a million.[11] The estimated number of deaths within 30 days of the initial diagnosis was 29 300 and the number of patients that experienced recurrence at least once was 83 000. Up to 20–35% of CDI patients would suffer from CDI recurrence, in which 45–65% of them would develop multiple recurrent episodes.[12,13] These recurrent episodes have posed a huge burden and clinical challenges in managing the disease.[9,13–15] Intelligent systems that assess CDI patients' severity at the early stage have important clinical implications in disease management, reduce risk infections' disease transmission, and improve clinical outcomes.[16]

With the changing epidemiology of CDI,[9,13–15] clinicians need immediate and reliable diagnostic tools to assess the disease severity and predict clinical outcomes. Therefore, a robust prediction system with statistical or machine learning models to identify the patients at a high risk of mortality or recurrence allows upfront planning of medical treatment to improve survival outcomes. A logistic regression (LR) model utilizing 2 065 patient data in two US academic institutions for predicting the inpatient mortality and other disease-related outcomes has been reported.[17] Random forest (RF), a machine learning algorithm utilizing the ensemble method, was applied in CDI recurrence prediction with 198 Caucasian patient data from two hospitals in 2014.[18] However, patient data acquired from some institutions lack generalizability when compared with administrative databases with a large sample size and patients from multicentered institutions. The Nationwide Inpatient Sample of the USA was studied to construct a CDI severity score using multivariate LR; however, neither clinical and treatment data nor laboratory results (e.g., white blood cell count) were presented in the database.[19] With more systematic data collection in the EHR system, statistical or machine learning models can utilize a more comprehensive set of patient features to generalize underlying clinical patterns and provide precise predictions.

Therefore, the aim of our study is to establish a prediction system based on a large, detailed and well-established EHR of over 15 000 CDI episodes to estimate CDI patients' mortality and recurrence outcomes.[10] We also compared the model accuracy of four various types of machine learning algorithms to regression analysis and evaluate the feature importance of the best-performing model to known clinical findings. The prediction system with the best-performing machine learning model is now available online for clinicians to assess and interpret the risk factors of CDI patients.

The study cohort is previously described.[10] In brief, CDI patient records diagnosed between 2006 and 2014 were obtained from 41 public hospitals in Hong Kong using the Clinical Data Analysis and Reporting System (CDARS), a well-established electronic database managed by the Hong Kong Hospital Authority comprising laboratory and clinical records covering over 90% of all inpatient services in the territory. Patients with a positive result on culture, toxin, or the molecular assay of a diarrhea stool sample were diagnosed as a CDI case. Mortality was defined according to patients' vital status within 30 days after CDI diagnosis, whereas recurrence was defined by a recurrent diarrhea stool specimen with a positive test result within 60 days after completion of CDI treatment.

The patient features for model developments are shown in **Table 1**. Features were grouped into patient demographics, admitting diagnosis, laboratory results, past surgical procedure, medication prescriptions, comorbid disease diagnoses, and clinical outcomes. Descriptive statistics are shown in Table 1. Continuous variables were presented as mean ± s.d. and compared using two-sample *t*-tests. Categorical variables were reported as *N* (%) and compared using Pearson's chi-square test (χ^2 test). The two-sample *t*-test aims at finding the statistically significant difference between two independent population means through comparing the two sample groups. χ^2 *test* of independence is used for proving the significant association between two categorical variables. The *p*-values of the tests are shown in Table 1.

Of the 15 168 patients in the CDI mortality study, 4,508 of them died due to CDI whereas 10 660 of them were alive after 30 days of diagnosis. Individuals with CDI-related mortality had an average age older than the living group (79.4 vs 72.6 years, $p < 0.001$). For other mortality risk factors reported in previous studies,[20–22] significant differences were obtained in four of them, including the presence of metastatic tumors (11.9% vs 4.7%, $p < 0.001$), healthcare-associated infection (96.0% vs 90.0%, $p < 0.001$), renal diseases (27.2% vs 21.7%, $p < 0.001$), and living in elderly homes (39.7% vs 27.9%, $p < 0.001$). Also, significantly higher mortality rates were observed for several laboratory parameters, such as a lower minimum albumin level (21.5 vs 27.1 g L^{-1}, $p < 0.001$), higher maximum creatinine level (231.2 vs 187.8 μmol L^{-1}, $p < 0.001$), higher maximum white blood cell counts (15.7 vs 12.1 kcells μL^{-1}, $p < 0.001$), higher maximum C-reactive protein (CRP) (9.8 vs 6.9 mg L^{-1}, $p < 0.001$), and creatinine rise over 150 μmol L^{-1} (18.2% vs 36.8%, $p < 0.001$).

We next analyzed the clinical parameters associated with disease recurrence. Among the 15 864 patients, 1 219 of them suffered CDI recurrence within 60 days whereas 14 645 of them did not. Healthcare-associated infection was significantly associated

Prof. M. Ip
Department of Microbiology
Faculty of Medicine
The Chinese University of Hong Kong
Hong Kong

Dr. L. Zhang, Dr. W. K. K. Wu
Department of Anaesthesia and Intensive Care
Faculty of Medicine
The Chinese University of Hong Kong
Hong Kong

Table 1. Patients characteristics and statistical analysis.

Patient features	Mortality [n = 15 168]			Recurrence [n = 15 864]		
	0 [Negative] [n = 10 660]	1 [Positive] [n = 4 508]	P Value	0 [Negative] [n = 14 645]	1 [Positive] [n = 1 219]	P Value
Demographics						
Age [mean ±SD]	72.57 ± 17.19	79.37 ± 13.20	<0.001	73.74 ± 16.87	75.88 ± 15.81	<0.001
Male [%]	5 064 [48%]	2 283 [51%]	<0.001	7 084 [48%]	599 [49%]	0.612
Female [%]	5 596 [53%]	2 225 [49%]	<0.001	7 561 [52%]	620 [51%]	0.612
Old age home [%]	2 969 [28%]	1 790 [40%]	<0.001	4 399 [30%]	397 [33%]	0.068
Recent hospitalization within 12 weeks [%]	8 353 [78%]	3 790 [84%]	<0.001	11 556 [79%]	1 038 [85%]	<0.001
Admitting diagnosis						
Onset within 48 h [%]	3 160 [30%]	858 [19%]	<0.001	3 915 [27%]	303 [25%]	0.157
Emergency Admission [%]	6 058 [57%]	2 648 [59%]	0.030	8 423 [58%]	646 [53%]	0.002
Episode times [mean ±SD]	1.23 ± 0.60	1.20 ± 0.55	0.001	1.21 ± 0.57	1.33 ± 0.71	<0.001
Severe IDSA [%]	4 309 [44%]	1 791 [44%]	0.911	5 762 [43%]	585 [50%]	<0.001
Healthcare associated [%]	9 587 [90%]	4 328 [96%]	<0.001	13 351 [91%]	1 165 [96%]	<0.001
Community associated [%]	642 [6%]	91 [2%]	<0.001	782 [5%]	23 [2%]	<0.001
Indetermined [%]	431 [4%]	89 [2%]	<0.001	512 [4%]	31 [3%]	0.089
Laboratory results						
Maximum white blood cell counts [mean ± SD]	12.06 ± 10.06	15.70 ± 11.73	<0.001	13.02 ± 10.65	13.57 ± 10.58	0.120
Minimum albumin level [mean ± SD]	27.08 ± 6.69	21.55 ± 6.02	<0.001	25.73 ± 7.10	24.64 ± 6.34	<0.001
Maximum creatinine level [mean ± SD]	187.80 ± 251.78	231.21 ± 266.50	<0.001	198.73 ± 257.16	188.94 ± 241.29	0.216
Maximum CRP [mean ± SD]	6.90 ± 7.02	9.83 ± 7.53	<0.001	7.72 ± 7.29	8.35 ± 7.37	0.131
Creatinine rise above 150 [%]	3 245 [37%]	654 [18%]	<0.001	3 633 [30%]	459 [45%]	<0.001
White blood cell 15 [%]	1 935 [23%]	1 407 [41%]	<0.001	3 155 [28%]	289 [29%]	0.356
Past surgical procedure						
Therapy [%]	6 895 [65%]	2 975 [66%]	0.122	8 958 [61%]	1 210 [99%]	<0.001
Surgery intervention [%]	1 483 [14%]	445 [10%]	<0.001	1 946 [13%]	158 [13%]	0.789
Colectomy [%]	42 [0%]	13 [0%]	0.375	59 [0%]	1 [0%]	0.088
Medication prescriptions						
Antibiotics [%]	9 650 [91%]	4 248 [94%]	<0.001	13 261 [91%]	1 175 [96%]	<0.001
Penicillins [%]	7 709 [72%]	3 743 [83%]	<0.001	10 850 [74%]	972 [80%]	<0.001
Benzylpenicillin and phenoxymethylpenicillin [%]	45 [0%]	13 [0%]	0.252	54 [0%]	7 [1%]	0.232
Penicillinase resistant penicillins [%]	821 [8%]	410 [9%]	0.004	1 181 [8%]	98 [8%]	0.995
Broad spectrum pencillins [%]	6 884 [65%]	3 413 [76%]	<0.001	9 756 [67%]	860 [71%]	0.005
Antipseudomonal penicillins extended spectrum [%]	2 748 [26%]	1 465 [33%]	<0.001	3 965 [27%]	371 [30%]	0.012
Cephalosporins carbapenems and other beta-lactams [%]	5 175 [49%]	2 402 [53%]	<0.001	7 198 [49%]	684 [56%]	<0.001
Cephalosporins [%]	3 305 [31%]	1 648 [37%]	<0.001	4 461 [31%]	492 [40%]	<0.001
Carbapenems [%]	1 465 [14%]	655 [15%]	0.209	1 922 [13%]	198 [16%]	0.003
Tetracyclines [%]	141 [1%]	76 [2%]	0.084	208 [1%]	16 [1%]	0.886
Aminoglycosides [%]	1 120 [11%]	434 [10%]	0.108	1 495 [10%]	135 [11%]	0.351
Macrolides [%]	999 [9%]	485 [11%]	0.009	1 432 [10%]	98 [8%]	0.051
Clindamycin [%]	93 [1%]	39 [1%]	0.995	126 [1%]	10 [1%]	1.000
Anti-tuberculosis drugs [%]	405 [4%]	147 [3%]	0.108	538 [4%]	57 [5%]	0.081
Anti-leprotic drugs [%]	46 [0%]	10 [0%]	0.061	58 [0%]	2 [0%]	0.326
Metronidazole and tinidazole [%]	2 338 [22%]	1 003 [22%]	0.669	3 054 [21%]	398 [33%]	<0.001
Quinolones [%]	3 751 [35%]	1 646 [37%]	0.123	5 067 [35%]	521 [43%]	<0.001
H2 receptor antagonists [%]	4 968 [47%]	2 207 [49%]	0.008	6 854 [47%]	607 [50%]	0.045
Prostaglandin analogues [%]	5 [0%]	6 [0%]	0.096	8 [0%]	3 [0%]	0.047
Proton-pump inhibitors [%]	4 843 [45%]	2 474 [55%]	<0.001	6 869 [47%]	637 [52%]	<0.001

Table 1. Continued.

Patient features	Mortality [$n = 15\,168$]			Recurrence [$n = 15\,864$]		
	0 [Negative] [$n = 10\,660$]	1 [Positive] [$n = 4\,508$]	*P* Value	0 [Negative] [$n = 14\,645$]	1 [Positive] [$n = 1\,219$]	*P* Value
Laxatives [%]	5 285 [50%]	2 635 [59%]	<0.001	7442 [51%]	685 [56%]	<0.001
Immunosuppressants [%]	2 751 [26%]	1 131 [25%]	0.360	3 782 [26%]	283 [23%]	0.048
Glucocoticoid therapy [%]	2 131 [20%]	962 [21%]	0.061	2 982 [20%]	233 [19%]	0.317
Alkylating drugs [%]	442 [4%]	89 [2%]	<0.001	541 [4%]	39 [3%]	0.425
Anthracyclines and other cytotoxic antibiotics [%]	283 [3%]	61 [1%]	<0.001	348 [2%]	24 [2%]	.428
Antimetabolites [%]	684 [6%]	159 [4%]	<0.001	843 [6%]	58 [5%]	0.159
Vinca alkaloids and etoposide [%]	309 [3%]	93 [2%]	0.004	403 [3%]	29 [2%]	0.517
Other antineoplastic drugs [%]	406 [4%]	169 [4%]	0.887	555 [4%]	36 [3%]	0.159
Antiproliferative immunosuppressants [%]	304 [3%]	73 [2%]	<0.001	395 [3%]	21 [2%]	0.044
Corticosteriods and other immunosuppressants [%]	332 [3%]	69 [2%]	<0.001	427 [3%]	14 [1%]	<0.001
Anti-lymphocyte monoclonal antibodies [%]	94 [1%]	35 [1%]	0.559	126 [1%]	13 [1%]	0.424
Other immunomodulating drugs [%]	48 [1%]	10 [0%]	0.047	53 [0%]	5 [0%]	0.802
Drugs that suppress the rheumatic disease process [%]	111 [1%]	31 [1%]	0.044	141 [1%]	7 [1%]	0.214
Co-morbid disease diagnoses						
Refractory disease [%]	1 556 [15%]	616 [14%]	0.070	1 905 [13%]	320 [26%]	<0.001
Myocardial infarction [%]	924 [9%]	548 [12%]	<0.001	1 353 [9%]	133 [11%]	0.056
Congestive heart failure [%]	2 275 [21%]	1 312 [29%]	<0.001	3 338 [23%]	285 [23%]	0.646
Peripheral vascular disease [%]	577 [5%]	284 [6%]	0.031	802 [6%]	78 [6%]	0.170
Cerebrovascular accident [%]	3 446 [32%]	1 596 [35%]	<0.001	4 622 [32%]	522 [43%]	<0.001
Dementia [%]	1 030 [10%]	521 [12%]	<0.001	1 419 [10%]	147 [12%]	0.008
Chronic peritoneal dialysis [%]	1 608 [15%]	747 [17%]	0.022	2 206 [15%]	182 [15%]	0.930
Rheumatic disease [%]	134 [1%]	47 [1%]	0.288	175 [1%]	13 [1%]	0.784
Peptic ulcer disease [%]	1 316 [12%]	646 [14%]	0.001	1 842 [13%]	166 [14%]	0.302
Mild liver disease [%]	1 454 [14%]	599 [13%]	0.567	1 999 [14%]	149 [12%]	0.178
Diabetes without chronic complication [%]	1 952 [18%]	863 [19%]	0.235	2 685 [18%]	234 [19%]	0.466
Diabetes with chronic complication [%]	1 145 [11%]	518 [12%]	0.181	1 550 [11%]	135 [11%]	0.597
Hemiplegia or paraplegia [%]	891 [8%]	358 [8%]	0.401	1 143 [8%]	137 [11%]	<0.001
Renal disease [%]	2 310 [22%]	1 226 [27%]	<0.001	3 296 [23%]	298 [24%]	0.125
Blood or non-metastatic solid tumor [%]	2 541 [24%]	1 299 [29%]	<0.001	3,717 [25%]	269 [22%]	0.010
Moderate or severe liver disease [%]	220 [2%]	105 [2%]	0.298	317 [2%]	20 [2%]	0.257
Metastatic solid tumor [%]	501 [5%]	537 [12%]	<0.001	992 [7%]	55 [5%]	0.002
Acquired immune deficiency syndrome [%]	58 [1%]	13 [0%]	0.040	75 [1%]	3 [0%]	0.284
Inflammatory bowel disease [%]	113 [1%]	11 [0%]	<0.001	128 [1%]	10 [1%]	1.000
Crohn disease [%]	39 [0%]	4 [0%]	0.002	51 [0%]	2 [0%]	0.436
Ulcerative colitis [%]	83 [1%]	7 [0%]	<0.001	86 [1%]	8 [1%]	0.698

with a higher CDI recurrence rate (95.6% vs 91.4%, $p < .001$), as with patients with creatinine rise over 150 μmol L^{-1} (45.0% vs 30.3%, $p < 0.001$). Significantly higher recurrent rates were observed for patients with lower minimum albumin levels (24.6 vs 25.7 g L^{-1}, $p < 0.0001$), higher age (75.9 vs 73.7 years, $p < 0.001$), and the number of past CDI episodes (1.3 vs 1.2, $p < 0.001$).

To improve the generalization for CDI mortality and recurrence prediction, all data were divided into training and test sets according to patients' healthcare institutions. This external validation method is preferred over random splitting on training data and testing data because factors such as clinical settings or populations may vary according to hospitals. The external validation method not only preserves the characteristics from distinct institutions but also can evaluate the model accuracy and reliability when the models are transported to other clinical environments and assessing different populations.[23–25] The training set contained 76% of the samples from 33 hospitals (11 470 samples in mortality prediction and 12 024 samples in recurrence prediction), and the test set contained the remaining 24% from 8 hospitals (3 698 samples in mortality prediction and 3 840 samples in recurrence prediction).

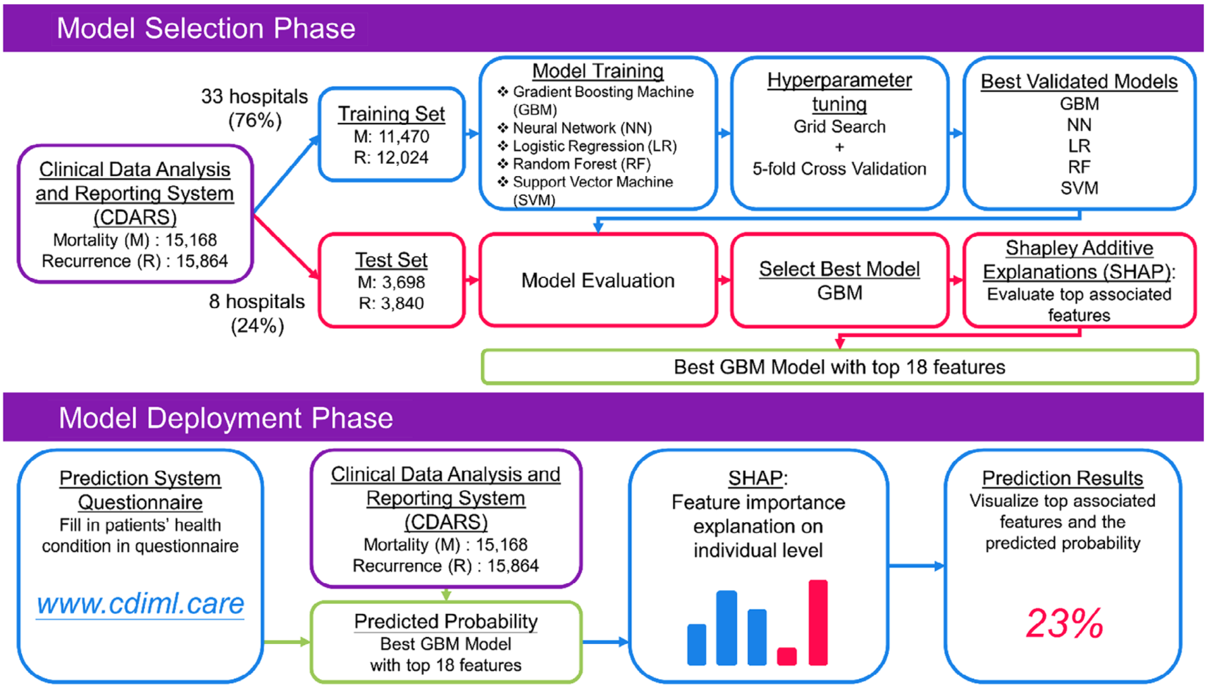
Model Selection Phase
33 hospitals (76%)
Clinical Data Analysis and Reporting System (CDARS)
Mortality (M) : 15,168
Recurrence (R) : 15,864
Training Set
M: 11,470
R: 12,024
Model Training
❖ Gradient Boosting Machine (GBM)
❖ Neural Network (NN)
❖ Logistic Regression (LR)
❖ Random Forest (RF)
❖ Support Vector Machine (SVM)
Hyperparameter tuning
Grid Search
+
5-fold Cross Validation
Best Validated Models
GBM
NN
LR
RF
SVM
8 hospitals (24%)
Test Set
M: 3,698
R: 3,840
Model Evaluation
Select Best Model
GBM
Shapley Additive Explanations (SHAP):
Evaluate top associated features
Best GBM Model with top 18 features
Model Deployment Phase
Prediction System Questionnaire
Fill in patients' health condition in questionnaire
www.cdiml.care
Clinical Data Analysis and Reporting System (CDARS)
Mortality (M) : 15,168
Recurrence (R) : 15,864
Predicted Probability
Best GBM Model with top 18 features
SHAP:
Feature importance explanation on individual level
Prediction Results
Visualize top associated features and the predicted probability
23%

Figure 1. Workflow for machine learning model selection and deployment. In the model selection phase, grid search cross validation was applied to tune the optimal hyperparameter for each algorithm. The optimal model in each algorithm was evaluated in the test set to determine the best-performing algorithm for the dataset. SHAP analysis was applied to identify the top associated features. In model deployment phase, the model was retrained with top 18 features identified in the SHAP analysis and integrated to an open-access prediction system at https://www.cdiml.care/.

The overall workflow of model selection and development is shown in **Figure 1**. Five prognostic models on CDI mortality and recurrence were developed using the statistical LR model and four various types of machine learning algorithms, including support vector machines (SVM), GBM, RF, and neural network (NN). LR and SVM are models that maximize the n-dimensional feature distance to distinctly classify the data points. RF and GBM, which recursively split the features in a top-down induction manner to distinguish between classes, utilize bootstrap aggregating and boosting methods correspondingly to convert weak decision tree classifiers into strong classifiers. NN mimics the human brain for pattern recognition and utilized a feedback mechanism to learn the intrinsic weight by minimize classification error. CDI mortality and recurrence score systems were generally developed with statistical LR,[17,19,26] considering its ease of interpretation in model coefficients; nonetheless, the other four machine learning algorithms were adopted to handle more complicated problems such as detecting *C. difficile* toxins in stools[27] or modeling immunoregulatory therapeutics for treating CDI.[28] To avoid selection bias and reduce overfitting, the optimal hyperparameters for each model were determined by grid search and validated through fivefold cross validation.[29] The final model was determined by the hyperparameters producing the lowest error. To evaluate each model's discrimination power, the area under receiver operating characteristics (ROC) curves (AUC) was measured. The model with the highest AUC was chosen to develop into an open-access prediction system.

The ROC curves for the five predictive models in CDI mortality and recurrence prediction are shown in **Figure** 2A,B. The boosting-based machine learning algorithm GBM (Mortality AUC: 0.7878; Recurrence AUC: 0.7076) outperformed statistical-based LR models (Mortality AUC: 0.7573; Recurrence AUC: 0.6927) or other types of machine learning algorithms. The outstanding results of the boosting-based algorithm GBM over the other four algorithms may be attributed to two reasons. First, boosting,[30] a method which combines weak classifiers into a final strong classifier and increases the previous misclassified data weights for future training, improves the capability of handling the imbalanced characteristics in CDI dataset.[31] Second, GBM can well generalize unseen data because it tends to reduce errors (bias and variance) during the training process. A shallow tree in a weak classifier has a high bias but low variance. GBM builds the weak classifiers sequentially so that the error of prior classifiers can be modeled by posterior classifiers to reduce the overall bias.[32] These properties make GBM a more robust model for CDI mortality and recurrence prediction.

Understanding the top associated patient features, which lead to specific prediction in machine learning models, is a key challenge in machine learning. Complex models such as GBM and NN may provide significantly improved accuracy, however, with more complication in interpreting the hidden reasons. Therefore, we adopted SHAP,[33] which applies a technique in game theory to quantify the contribution of each player in a collaborative game, to represent the feature importance distributed in each classifier of the model. The feature importance of the machine learning model facilitates us to interpret the relevant features for output prediction and exclude the irrelevant features to reduce model complexity.[34]

The feature importance of each classifier in the GBM model to predict CDI mortality and recurrence is shown in **Figure 3**.

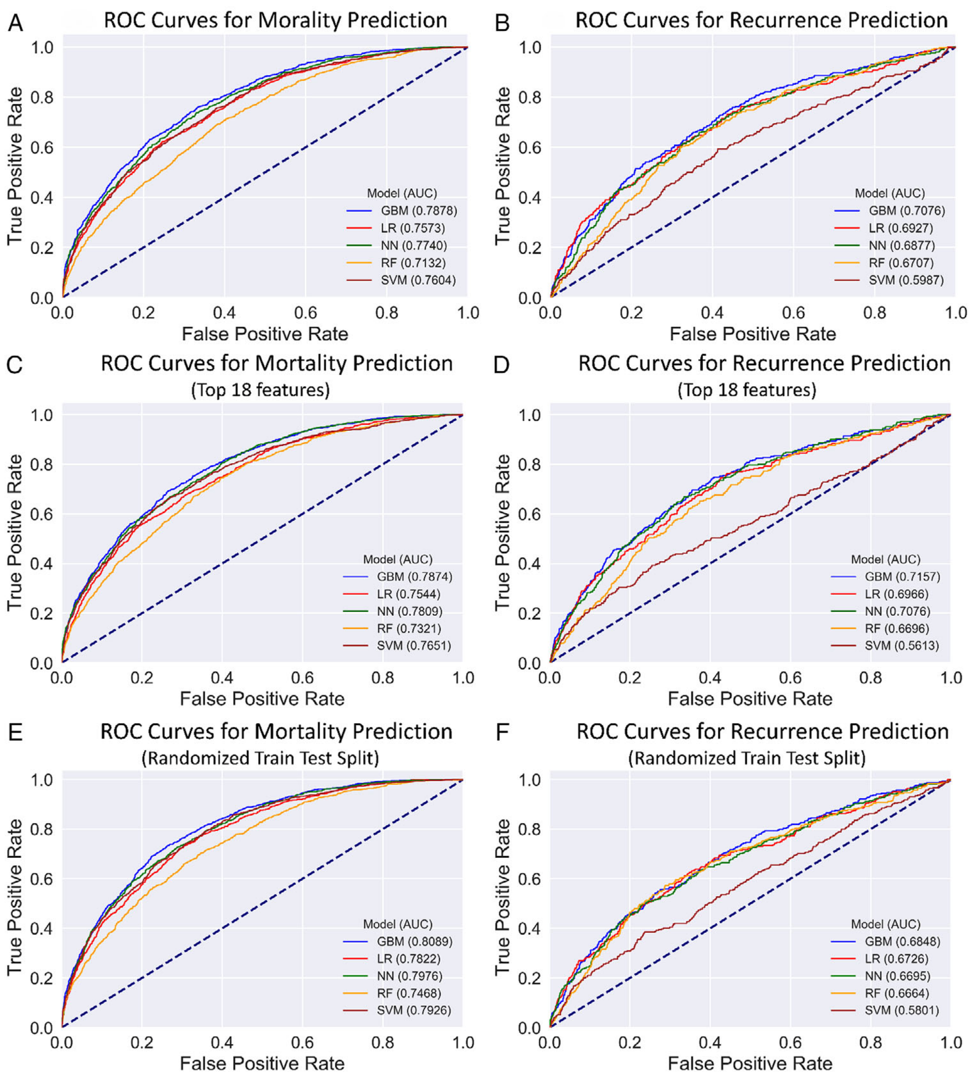

Figure 2. ROC curves of four machine learning models and the LR model in the test set. A) Mortality prediction and B) recurrence prediction. C) Mortality prediction and D) recurrence prediction of models trained with top 18 features only. E) Mortality prediction and F) recurrence prediction without train test data splitting corresponding to patients' healthcare institution.

The features ranked at the top imply that the models choose these features more frequently to correctly classify patients. The color and the distribution of samples reveal the relationship between risk prediction and feature value. The low value of the minimum albumin level, long hospitalization, absence of creatinine level over 150 μmol L^{-1} and aging increase the mortality probability of patients. Number of days hospitalized, creatinine level rise 150 μmol L^{-1} and metronidazole prescription before are the three leading features to predict CDI recurrence.

To evaluate with existing clinical findings, risk factors associated with CDI mortality and recurrence from previous studies were found.[10,20,35–37] Significant associations were found between CDI mortality and several patient factors, including older age, reduced albumin levels, higher creatinine levels, and higher white blood cell counts. In Figure 3A, the feature importance of GBM mortality model shows that the abovementioned factors are also shown as the top-associated features for accurate prediction, which is consistent with existing clinical findings. For disease recurrence, former studies identified several risk factors including aging, long-term hospitalization, healthcare-associated infection, severe CDI, and hospitalization in common isolation units.[10,37–41] Aging and hospitalized days over 30 days are shown as top predictive recurrence risk factors in Figure 3B. The use of machine learning algorithms suggested other key risk factors, for instance, the creatinine rise above 150 μmol L^{-1} previous usage of nitroimidazole e.g., metronidazole, and refractory disease.

To facilitate CDI risk assessments, the best-performed CDI mortality and recurrence predictive models were validated and released online as a platform for worldwide access at: https://www.cdiml.care/. Clinicians may fill in the questionnaire for individuals or upload comma-separated values files consisting multiple patients' data in our web application to obtain prediction results. The questionnaire was designed to include the top 18

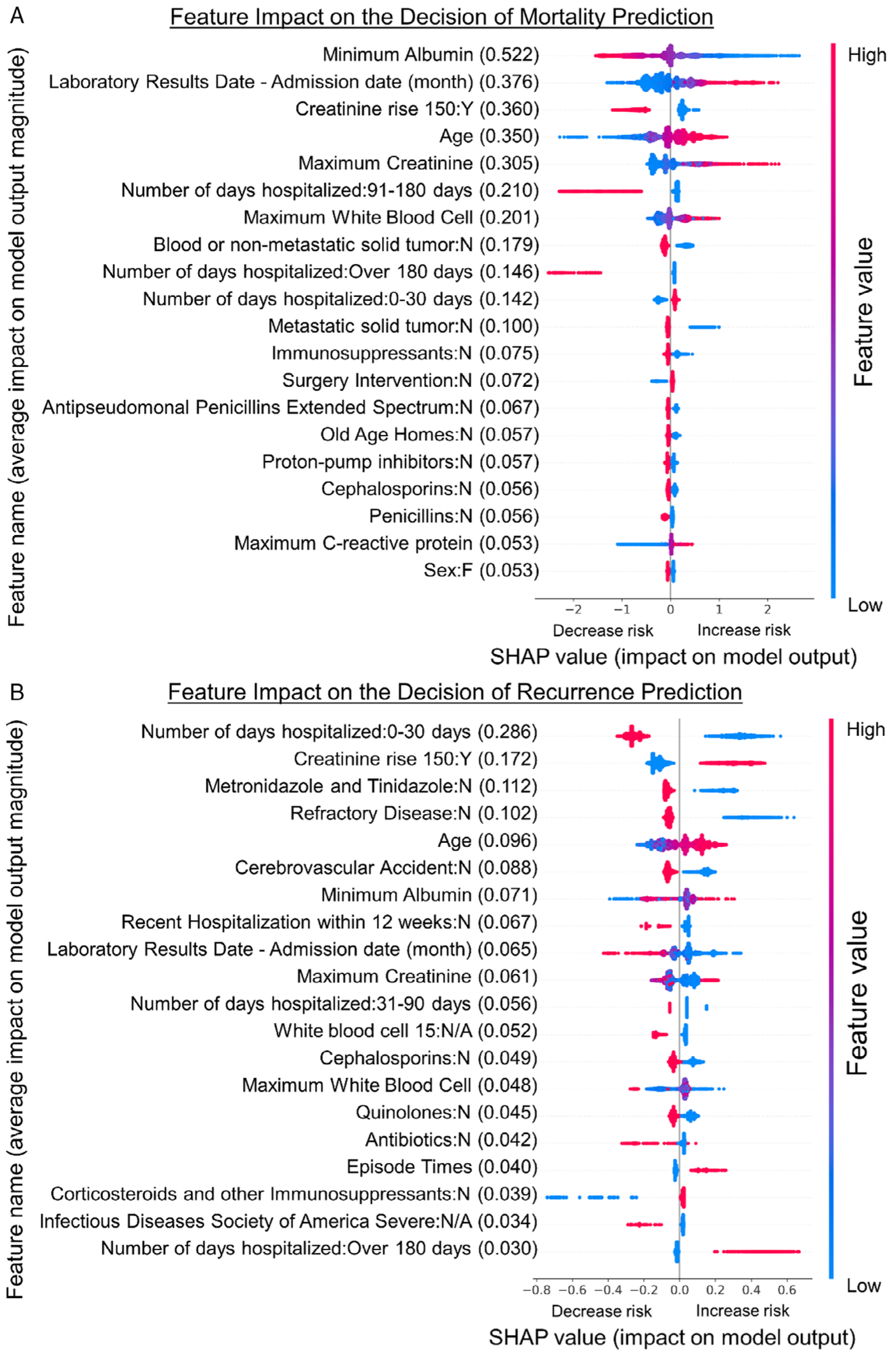

Figure 3. Top 20 feature importance ranked according to the prediction. The red indicates high feature values and the blue indicates low ones. Positive SHAP value shows how the positive predicted class relates to the feature value. Note that one categorical feature 'number of days hospitalized' contains three cases (0–30 days, 91–180 days, over 180 days) ranked in the top 20 feature importance.

most relevant features as input to facilitate efficient clinical assessment. To assess whether the GBM would outperform the other 4 algorithms using the top 18 features out of all 79 features, all 5 models were trained under the reduced feature setting for mortality and recurrence prediction. The GBM achieved an AUC of 0.7874 in mortality prediction and 0.7157 in recurrence prediction. Figure 2C,D shows the AUC in the top 18 feature setting. Comparing the AUC of the GBM model with originally 79 features and the GBM model with the top 18 associated features, the AUC is from 0.7878 to 0.7874 in mortality prediction and 0.7076 to 0.7157 in recurrence prediction. The GBM model trained with top associated features yields comparable or even more excellent prediction results in contrast to the full-feature GBM model.

Risk factors explanation on an individual level was incorporated on the open-access platform to complement the prediction probability to demonstrate the reason why the machine learning model produces certain predictions. The individual-level explanation is more specific for clinicians to assess which factors influence the patient most rather than the entire population. **Figure 4**A shows the risk factor explanation of a patient who was predicted to have a high mortality probability of 0.81. The risk factors increasing or decreasing the predicted mortality risk can be explained by patient features listed on the red or blue bars, respectively: the red bars indicate the risk factors pushing the mortality prediction higher and the blue bars indicate the risk lowering the mortality prediction. Features are sorted according to the magnitude toward probability prediction. Examining the risk factor explanation of the patient, 8.52×10^6 cells mL^{-1} of white blood cell count is within a normal range and, therefore, shows up as a factor to reduce the mortality probability. However, the levels of albumin and creatinine are beyond the standard range, thus increasing mortality probability. A patient who was predicted to have low mortality probability of 0.03 is shown in Figure 4B. The predicted low mortality risk can be explained by the normal albumin level (34–54 g L^{-1}), middle age, and normal white blood cell count ($\leq$15 kcells μL^{-1}). With the individual-level risk explanation, professionals can have a clearer understanding of the relationship on which input patient features increase or decrease the predicted mortality probability, rather than a predicted value from the machine learning models.

There were several limitations present in this study. First, these experiments were conducted in retrospect. In addition, all the patient data in this study were from Hong Kong. Sample bias may exist due to regional restrictions. As only 79 prediction features were included in this study, some associated factors may not be covered in this investigation. Furthermore, the clinical benefit of the machine learning approach (over conventional risk prediction) will need to be proven in prospective clinical studies. A randomized controlled trial can demonstrate improved patient outcomes using this algorithm, which may predict high-risk patients not identified by conventional clinical parameters.

The performances of our model and previously reported models are evaluated in depth. In mortality prediction, Kulaylat, Audrey S, et al.[17] developed a multivariable LR model with an AUC of 0.82 and Z. Kassam et al.[19] reported a severity score model based on multivariate LR to achieve an AUC of 0.77. The AUC of our best GBM model is 0.7878, which is comparable with their performance. It is worth mentioning that in a study by Kulaylat et al.[17] the eosinophil count acted as a statistically significant variable in assessing the mortality risk, which could be a potential factor where the models cannot attain a higher AUC in our study cohort. In recurrence prediction, we observed a noticeable AUC difference between our study and the study by LaBarbera, Francis D, et al.[18] One potential reason is that their study consisted a majority of Caucasian patients while our study is Hong Kong based, where demographic variation may exist. In addition, RF algorithm is included in our study to investigate the predictive power of different machine learning algorithms.

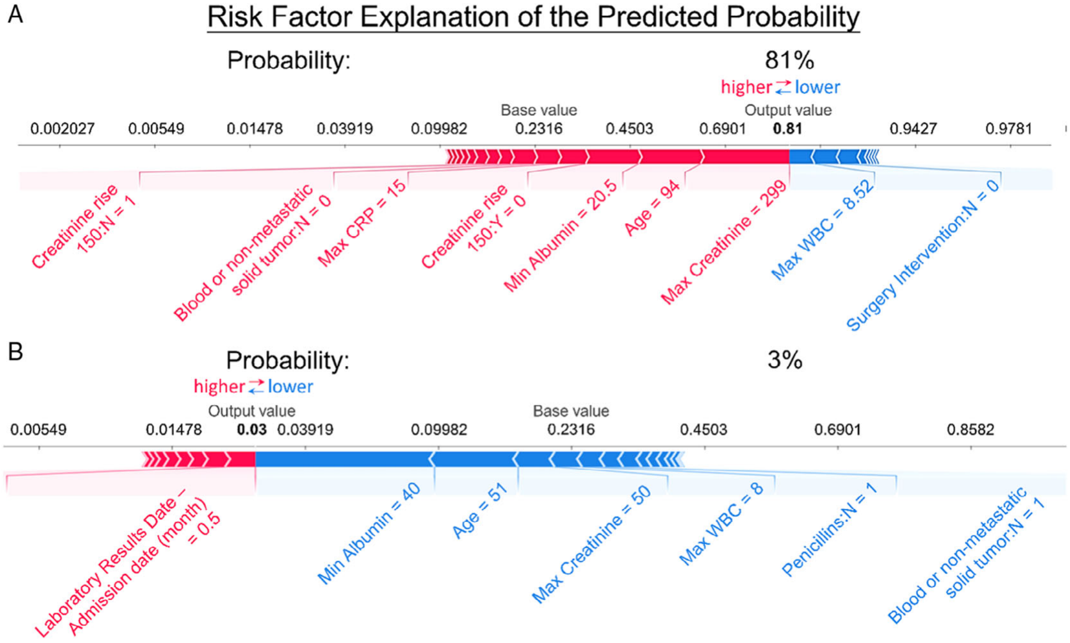

Figure 4. Risk factor explanation of the predicted probability by GBM. The red bars indicate the risk factors pushing the mortality prediction higher and the blue bars indicate the risk lowering the mortality prediction. A) A patient predicted to have high mortality risk. B) A patient predicted to have low mortality risk.

The best AUC of the RF algorithm can only achieve 0.6707, which is lower than our best GBM model.

In conclusion, CDI mortality and recurrence predictive models were set up using different statistical and machine learning algorithms, the boosting-based algorithm GBM achieved the best performances in both mortality and recurrence prediction. The best-performed predictive models were available online for use. With the changing epidemiology of incipient and recurrent CDI,[10,11,42] such mortality and recurrence prognosis models hopefully could act as reference for doctors during treatment planning. Current treatment regimen recommended by the Infectious Diseases Society of America (IDSA) to treat initial CDI is the prescription of vancomycin or fidaxomicin, whereas high-dose vancomycin is suggested on fulminant CDI patients. Treatment intensification can be chosen at the early stage if patients are predicted to have worse clinical outcomes or heightened mortality. The recommended treatment for recurrent CDI is the tapered course of vancomycin, fecal microbiota transplantation, or other pharmacotherapies. Patients estimated to have high recurrence risk might benefit from upfront fecal microbiota transplantation or other pharmacotherapies before another recurrent episode. This may facilitate the early choice of appropriate treatment to reduce CDI mortality and recurrence, hence, relieving the healthcare burden. Intricate modeling and algorithms deterring clinicians from incorporating the machine learning model in patient care can be ameliorated with our intuitively accessible prediction website. The predicted probabilities and the individual level risk factors explanations can be conveniently acquired based on the entering clinical parameters. In addition, the top mortality-associated and recurrence-associated factors were uncovered through the machine learning algorithm. This could serve as a basis for further CDI pathology and clinical studies.

Experimental Section

Clinical Data Test: The statistical and machine learning model training were performed using Python 3.6.9. The data were divided into the training set (33 hospitals, 76% records) and test set (8 hospitals, 24% records). Data preprocessing steps were applied to handle missing data and categorical data. Numeric fields were first imputed by the mean value of the training samples to handle missing values and subsequently normalized through minimum–maximum normalization to avoid large varying feature values. For categorical variables, one-hot encoding was applied to represent the data structure for computation. The number of features increased from 79 to 222 during the data preprocessing steps. Randomized grid search cross validations were applied to tune the optimal hyperparameters for each learning algorithm. The randomized grid search techniques sample hyperparameter candidates from the target parameter sets and recursively validate them through the five-fold cross-validation process. For a total of 48 times, fivefold cross validations were carried out for each algorithm to explore the optimal hyperparameters. The optimal hyperparameter sets were determined by the highest mean validation score, which was obtained through averaging the fivefold cross-validation. To compare the model accuracy of different algorithms, the model with optimal hyperparameters of each algorithm were chosen and evaluated on the test set. In mortality prediction, the best-performed model was GBM with hyperparameter 190 estimators, 180 leaves, subsample ratio of 1, maximum tree depth of 2, and learning rate of 0.2, which achieved 0.811 mean validation score and 0.7878 test score. The best-performed model in recurrence prediction was GBM with 200 estimators, 180 leaves, subsample ratio of 1, maximum tree depth of 2, and learning rate of 0.05. The mean validation score was 0.693 and the test score was 0.7076. With a setting of randomized train test split, an identical set of model training and testing was conducted without splitting according to patients' healthcare institution. GBM also outperformed the other four algorithms with an AUC of 0.8089 in mortality prediction and 0.6848 in recurrence prediction. The ROC curves are shown on Figure 2E,F. The algorithms with the highest test accuracy were selected as the best-performing models and further analyzed with SHAP. The importance of each feature was ranked and is shown in Figure 3. An open-access prediction system was established based on the best-performed models with the top 18 most associated features.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Y.L.N. and M.C.K.L. contributed equally to this work. This work was supported by Research Grants Council (reference numbers: 17205919, 17206818, 17202317, and 24103516), National Natural Science Foundation of China (reference number: 81922082), Innovation and Technology Commission (reference numbers: UIM/353, SST/106/20GP), and Health and Medical Research Fund (reference number: CID-CUHK-C). *The Supporting Information of this article can be found here:* https://doi.org/10.22541/au.160340489.91686617/v2.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

Clostridioides difficile, diarrhea, infections, machine learning, pseudomembranous colitis

Received: August 12, 2020
Revised: October 19, 2020
Published online: November 16, 2020

[1] D. Bzdok, N. Altman, M. Krzywinski, *Nat. Methods* **2018**, *15*, 233.
[2] A. T. Hale, D. P. Stonko, A. Brown, J. Lim, D. J. Voce, S. R. Gannon, T. M. Le, C. N. Shannon, *Neurosurg. Focus* **2018**, *45*, E2.
[3] G. Liu, Y. Xu, X. Wang, X. Zhuang, H. Liang, Y. Xi, F. Lin, L. Pan, T. Zeng, H. Li, X. Cao, G. Zhao, H. Xia, *Sci. Rep.* **2017**, *7*, 1.
[4] J. L. Marcus, L. B. Hurley, D. S. Krakower, S. Alexeeff, M. J. Silverberg, J. E. Volk, *Lancet HIV* **2019**, *6*, e688
[5] Y. Yang, M. A. Bevan, B. Li, *Adv. Intell. Syst.* **2020**, https://doi.org/10.1002/aisy.201900106.
[6] K. Chin, T. Hellebrekers, C. Majidi, *Adv. Intell. Syst.* **2020**, *2*, 1900171.
[7] J. D. L. Ho, K. H. Lee, W. L. Tang, K. M. Hui, K. Althoefer, J. Lam, K. W. Kwok, *Adv. Robot.* **2018**, *32*, 1168.
[8] G. Fang, X. Wang, K. Wang, K. H. Lee, J. D. L. Ho, H. C. Fu, D. K. C. Fu, K. W. Kwok, *IEEE Robot. Autom. Lett.* **2019**, *4*, 1194.
[9] G. K. B. Ong, T. J. Reidy, M. D. Huk, F. R. Lane, *Am. J. Surg.* **2017**, *213*, 565.
[10] J. Ho, R. Z. W. Dai, T. N. Y. Kwong, X. Wang, L. Zhang, M. Ip, R. Chan, P. M. K. Hawkey, K. L. Y. Lam, M. C. S. Wong, G. Tse, M. T. V. Chan, F. K. L. Chan, J. Yu, S. C. Ng, N. Lee, J. C. Y. Wu, J. J. Y. Sung, W. K. K. Wu, S. H. Wong, *Emerg. Infect. Dis.* **2017**, *23*, 1671.

[11] N. Banaei, V. Anikst, L. F. Schroeder, *N. Engl. J. Med.* **2015**, *372*, 2368.
[12] L. V. McFarland, G. W. Elmer, C. M. Surawicz, *Am. J. Gastroenterol.* **2002**, *97*, 1769.
[13] R. Rodrigues, G. E. Barber, A. N. Ananthakrishnan, *Infect. Control Hosp. Epidemiol.* **2017**, *38*, 196.
[14] J. A. O'Brien, B. J. Lahue, J. J. Caro, D. M. Davidson, *Infect. Control Hosp. Epidemiol.* **2007**, *28*, 1219.
[15] K. A. Mergenhagen, A. L. Wojciechowski, J. A. Paladino, *Pharmacoeconomics* **2014**, *32*, 639.
[16] M. Tavakoli, J. Carriere, A. Torabi, *Adv. Intell. Syst.* **2020**, *2*, 2000071.
[17] A. S. Kulaylat, E. L. Buonomo, K. W. Scully, C. S. Hollenbeak, H. Cook, W. A. Petri, D. B. Stewart, *JAMA Surg.* **2018**, *153*, 1127.
[18] F. D. LaBarbera, I. Nikiforov, A. Parvathenani, V. Pramil, S. Gorrepati, *J. Community Hosp. Intern. Med. Perspect.* **2015**, *5*, 26033.
[19] Z. Kassam, C. Cribb Fabersunne, M. B. Smith, E. J. Alm, G. G. Kaplan, G. C. Nguyen, A. N. Ananthakrishnan, *Aliment. Pharmacol. Ther.* **2016**, *43*, 725.
[20] E. R. Dubberke, A. M. Butler, K. A. Reske, D. Agniel, M. A. Olsen, G. D'Angelo, L. C. McDonald, V. J. Fraser, *Emerg. Infect. Dis.* **2008**, *14*, 1031.
[21] M. S. Rubin, L. E. Bodenstein, K. C. Kent, *Dis. Colon Rectum* **1995**, *38*, 350.
[22] B. A. Jobe, A. Grasley, K. E. Deveney, C. W. Deveney, B. C. Sheppard, *Am. J. Surg.* **1995**, *169*, 480.
[23] S. E. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. T. Donders, G. Derksen-Lubsen, D. E. Grobbee, K. G. M. Moons, *J. Clin. Epidemiol.* **2003**, *56*, 826.
[24] E. W. Steyerberg, S. E. Bleeker, H. A. Moll, D. E. Grobbee, K. G. M. Moons, *J. Clin. Epidemiol.* **2003**, *56*, 441.
[25] R. D. Riley, J. Ensor, K. I. E. Snell, T. P. A. Debray, D. G. Altman, K. G. M. Moons, G. S. Collins, *BMJ* **2016**, i13140.
[26] C. P. Kelly, *Clin. Microbiol. Infect.* **2012**, https://doi.org/10.1111/1469-0691.12046.
[27] S. K. Koya, S. Yurgelevic, M. Brusatori, C. Huang, L. N. Diebel, G. W. Auner, *J. Surg. Res.* **2019**, *2444*, 111.
[28] A. Leber, R. Hontecillas, V. Abedi, N. Tubau-Juni, V. Zoccoli-Rodriguez, C. Stewart, J. Bassaganya-Riera, *Artif. Intell. Med.* **2017**, *78*, 1.
[29] J. Lever, M. Krzywinski, N. Altman, *Nat. Methods* **2016**, *13*, 803.
[30] J. H. Friedman, *Ann. Stat.* **2001**, *29*, 1189.
[31] I. Brown, C. Mues, *Expert Syst. Appl.* **2012**, *39*, 3446.
[32] J. H. Friedman, *Comput. Stat. Data Anal.* **2002**, https://doi.org/10.1016/S0167-9473(01)00065-2.
[33] S. M. Lundberg, S.-I. Lee, in *Adv. Neural Inf. Process. Syst.* Vol *30*, (Eds.: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., Red Hook, NY **2017**, pp. 4765–4774.
[34] Y. Saeys, T. Abeel, Y. Van De Peer, presented at *Machine Learning and Knowledge Discovery in Databases*, Antwerp **2008**.
[35] W. J. Halabi, V. Q. Nguyen, J. C. Carmichael, A. Pigazzi, M. J. Stamos, S. Mills, *J. Am. Coll. Surg.* **2013**, *217*, 802.
[36] M. G. Bloomfield, J. C. Sherwin, E. Gkrania-Klotsas, *J. Hosp. Infect.* **2012**, *82*, 1.
[37] C. N. Abou Chakra, J. Pepin, S. Sirard, L. Valiquette, *PLoS One* **2014**, *9*, e98400.
[38] A. M. Seekatz, K. Rao, K. Santhosh, V. B. Young, *Genome Med.* **2016**, *8*, 1.
[39] D. A. García-Lecona, E. Garza-González, M. Padilla-Orozco, L. Mendoza-Flores, S. Flores-Treviño, S. Mendoza-Olazaran, A. Camacho-Ortiz, *Am. J. Infect. Control* **2018**, *46*, 103.
[40] M. P. Bauer, D. W. Notermans, B. H. Van Benthem, J. S. Brazier, M. H. Wilcox, M. Rupnik, D. L. Monnet, J. T. Van Dissel, E. J. Kuijper, *Lancet* **2011**, *377*, 9759.
[41] J. Pépin, L. Valiquette, M. E. Alary, P. Villemure, A. Pelletier, K. Forget, K. Pépin, D. Chouinard, *CMAJ* **2004**, *171*, 466.
[42] A. C. Clements, R. J. S. Magalhães, A. J. Tatem, D. L. Paterson, T. V. Riley, *Lancet Infect. Dis.* **2010**, *10*, 395.