Heliyon 11 (2025) e41350

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Predicting positive *Clostridioides difficile* test results using large-scale longitudinal data of demographics and medication history

Anh Pham [a], Robert El-Kareh [a,b], Frank Myers [b], Lucila Ohno-Machado [a,c], Tsung-Ting Kuo [a,c,d,*]

[a] *Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, CA, USA*
[b] *UCSD Health System, San Diego, CA, USA*
[c] *Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA*
[d] *Department of Surgery, Yale School of Medicine, New Haven, CT, USA*

ARTICLE INFO

Keywords:
Clostridioides difficile infection
Electronic health record
Machine learning
Decision support systems

ABSTRACT

Background: Clostridioides difficile infection is a major health threat. Healthcare institutions have strong medical and financial incentives to keep infections under control. Blanket testing at admission is in general not recommended, and current predictive models either used moderate sample sizes, over-inflated the number of covariates, or chose non-interpretable algorithms. We aim to develop models using patient data to predict positive *Clostridioides difficile* test results with discrimination performance, interpretable results, and a reasonable number of covariates that reflect health over a long-time span.

Materials and methods: We processed records from 157,493 University of California San Diego Health patients seen between January 01, 2016–July 03, 2019 with at least 6 months of medication history, excluding pregnant women, patients under 18, and prisoners. Three models (Logistic Regression, Random Forest, and Ensemble) were constructed using hyper-parameters selected through 10-fold cross-validation. Model performance was measured by the Area Under the Receiver Operating Characteristic Curve (AUROC). The model coefficients' odds ratios and p-values were calculated for the Logistic Regression model, as were Gini indices for Random Forest. Decision boundary analysis was conducted using pair-wise false positive and false negative cases each model would predict at a specific threshold.

Results: Logistic Regression, Random Forest, and Ensemble models yielded test AUROCs of 0.839, 0.851, and 0.866, respectively. Significant covariates that may affect risk include age, immunocompromised treatments, past antibiotic uses, and some medications for the gastrointestinal tract.

Conclusions: The models achieve high discrimination performance (AUROC >0.83). There is a general consensus among different analysis approaches regarding predictors that impact patients' chances of having a positive test, which may influence *Clostridioides difficile* risk, including features clinically proven to increase susceptibility. These human-interpretable models can help distinguish significant predictors that affect a patient's chance of testing positive, which may influence their *Clostridioides difficile* risk.

* Corresponding author. 100 College Street, New Haven, CT, USA.
E-mail address: tsung-ting.kuo@yale.edu (T.-T. Kuo).

https://doi.org/10.1016/j.heliyon.2024.e41350
Received 30 April 2024; Received in revised form 20 November 2024; Accepted 18 December 2024
Available online 18 December 2024

1. Introduction

1.1. Clostridioides difficile infection and difficulty in diagnosis

Clostridioides difficile infection (CDI) is caused by the *Clostridioides difficile* (*C. diff*) bacterium and may lead to colitis, pseudomembranous colitis, life-threatening diarrhea, and sepsis [1, 2], especially in older, antibiotic-treated patients in long-term care [3]. The Centers for Disease Control and Prevention (CDC) deems CDI a major health threat [4]; in 2017, there were an estimated 223900 cases among US hospitals, with 12800 deaths and an healthcare-associated (HA) cost of $1billion [5]. In addition to critical public health concerns, failure to meet the infection-specific standardized infection ratio (SIR) [6] for HA-CDI as set by CDC can damage the reputations of healthcare facilities and bring about consequential financial implications [7]. For example, the 2015–2017 HA-CDI rate at University of California San Diego (UCSD) Health was significantly worse than the 2015 national baseline SIR [8]. It is a crucial goal for institutions to reduce HA-CDI so as to assure quality of care, avoid financial penalties and provide superior patient outcomes.

As *C. diff* bacteria are highly resistant to standard hospital decontamination chemicals, highly capable to survive on surfaces disinfected with standard agents, and can form transmissible spores [9–11], it is appropriate to consider a proactive approach that is still respectful of current guidelines. One organization's experience was that admission screening on almost all patients prevented up to 62 % of expected cases, and resulting a gradual decrease of infections over time [12]. From a purely administrative standpoint, early screening may help distinguish true HA-CDI from community-onset incidents, thus improving the accuracy of CDC surveillance efforts, and easing performance stress on healthcare facilities. However, current guidelines do not recommend blanket testing at admission due to fear of overdiagnosis, and subsequently unnecessary antibiotic treatments that may lead to resistance [13]; rather, testing is administered at symptom onset [14]. This leads to tension between accurate diagnosis and timely intervention, given that carriers (those with colonization) who are asymptomatic at admission have higher risks of progressing to CDI [15], and that they may shed resistant spores and contribute to outbreaks [16, 17]. This practical need calls for innovative, data-driven screening methods that do not require upfront bio-sample testing, while still offering physicians with discriminative insights.

1.2. Current machine learning models to predict CDI and their drawbacks

To aid screening efforts, researchers have built stratification models to classify potential *C. diff* positive cases. Most models were built on data with (i) a modest number of patients (about 8000 to 36,000 patients [18, 19]), (ii) several thousand covariates (about 1800 to 5000 features, among those are binary encoding of variables not necessarily common among all patients [20]) leading to difficulty in human interpretation of artificially-inflated, high-dimensional vectors, and/or (iii) predictors that bias towards a narrow window of near-current, or current admission data (e.g., antibiotic use within 30 days prior to testing [21]) that may fail to capture the accumulated effect of microbiota-alteration medications over time [22].

Due to these limitations, we sought to construct CDI predictive models that address (i) large-scale sample size, by using a significantly large number of patients (i.e., 157,493 patients) from a real-world institution such as UCSD Health, (ii) high discrimination and interpretable results, employing a reasonable number of core demographics and medical history covariates (i.e., 104 features) that (iii) better reflect each patient's health history over a longitudinal time span (i.e., 3 years). By predicting positive test results which may differ from a true CDI diagnosis (i.e., it may present colonization), allowing for expert discretion from the care staff, our goal is to add to the physicians' decision-support toolbox to use with other testing strategies, finding a compromise between universal testing for asymptomatic carriers and waiting to test until the presence of severe symptoms.

1.3. Objective

Specifically, our predictive task is to predict the first event of a positive *C. diff* lab test given demographic and medication history up to one calendar day before the ordering date of the first positive test (Fig. 1). In most cases, this meant at least two calendar days before

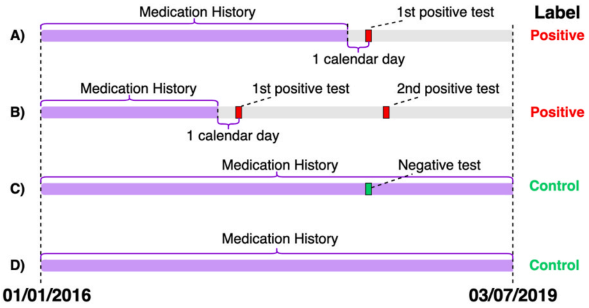

Fig. 1. Prediction task, illustrating examples of *Clostridioides difficile* Infection (CDI) Positive test result labels versus Control ones used in our study. The four cases for a patient are: **(A)** the patient has a single positive test during the study window, in which case they are labeled as a "Positive" case; **(B)** the patient has multiple positive tests, the first of which is also used as the prediction target of "Positive"; **(C)** the patient has a negative *C. diff* test and serves as a "Control" label; and **(D)** the patient never has any *C. diff* test and is a "Control" as well.

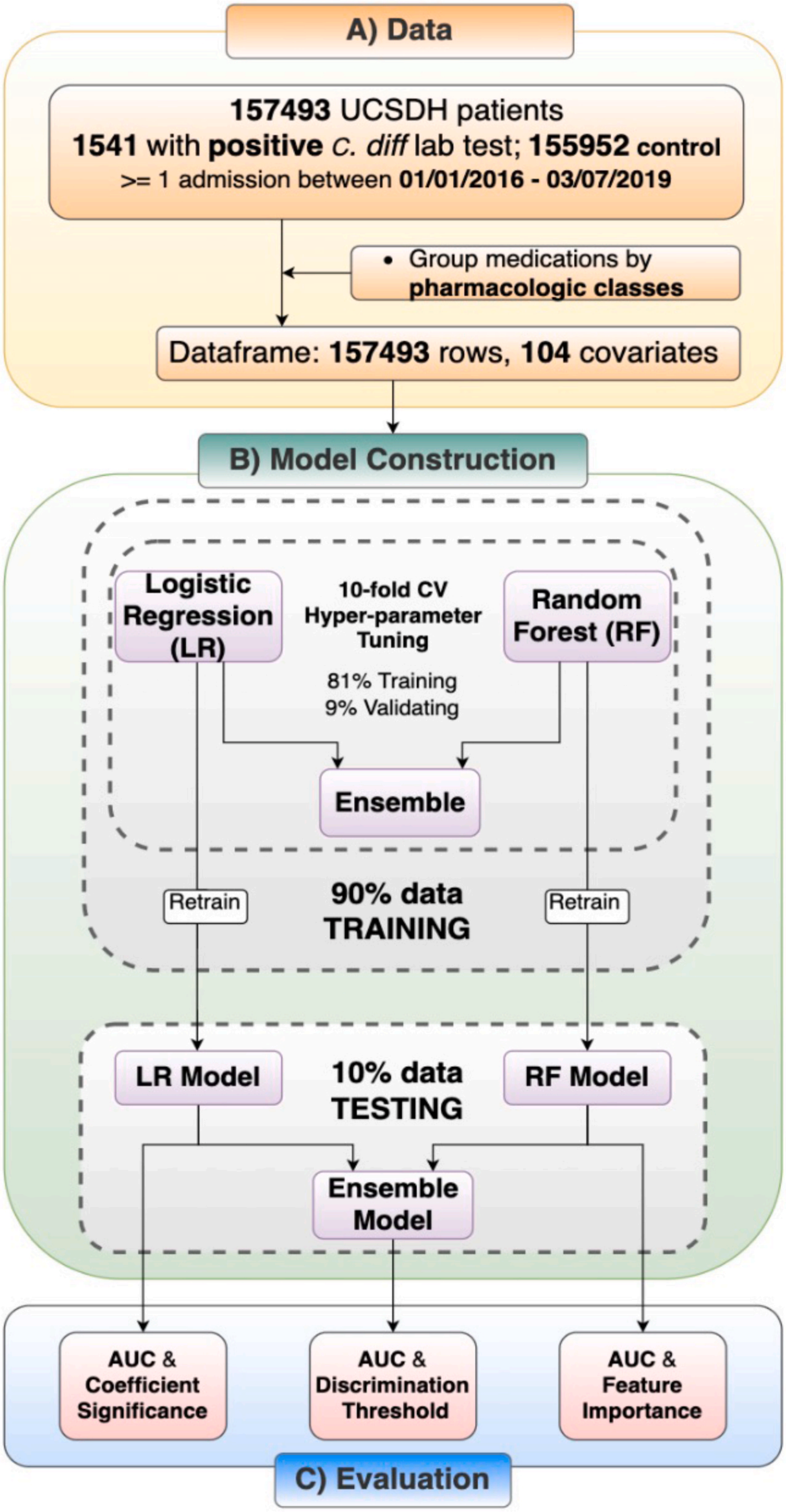
A) Data
157493 UCSDH patients
1541 with positive C. diff lab test; 155952 control
>= 1 admission between 01/01/2016 - 03/07/2019
• Group medications by pharmacologic classes
Dataframe: 157493 rows, 104 covariates
B) Model Construction
Logistic Regression (LR)
10-fold CV Hyper-parameter Tuning
Random Forest (RF)
81% Training
9% Validating
Ensemble
Retrain
90% data TRAINING
Retrain
LR Model
10% data TESTING
RF Model
Ensemble Model
AUC & Coefficient Significance
AUC & Discrimination Threshold
AUC & Feature Importance
C) Evaluation

Fig. 2. Study workflow. **(A)** Data: Demographics and medication history of 157493 patients with at least one admission to UCSD Health between 01/01/2016-03/07/2019 were vectorized. **(B)** Model Construction: Three algorithms (Logistic Regression, Random Forest, and Ensemble) were constructed. **(C)** Evaluation: The corresponding predictive performances of the models were recorded, and feature analysis carried out.

the result would be known, allowing physicians to proactively care for potential cases before their actual test confirmations. As adult carriers of toxicogenic *C. diff* are six times more likely to develop infections [15], this early prediction of colonization may make significant contributions to the fight against CDI.

2. Materials and methods

The overall workflow of our study is shown in Fig. 2. There are three main parts in our workflow: Data, Model Construction, and Evaluation. Each part will be introduced in the following subsections accordingly.

2.1. Data

The overall data preparing process is shown in Fig. 2A. We obtained data of 157,493 patients admitted between January 01, 2016 and July 03, 2019 (date approved by IRB# 190457CX) from the UCSD Health System. Our inclusion criteria included patients with at least one admission to UCSD Health facilities within the study window, whose data included at least six months of medication history. Our exclusion criteria included vulnerable subjects such as patients under the age of 18, pregnant women, and prisoners, as well as those who did not have age information. Among included patients, there were 1541 (1 %) CDI Positive test result labels defined in Fig. 1.

We chose a core number of covariates that encompass demographic information and medication history, which naturally reflect a patient's pathophysiology and health development over time. The dataset contained 104 covariates, including (a) demographic details of race, age, and gender as inherited from the UCSD Health System metadata schema (10 demographic predictors), and (b) past medication history in the form of tabulated instances used within the time frame (94 medication predictors). Specifically, medications were grouped from 4420 individual names to 94 respective pharmacologic classes, to ensure computational feasibility, reduce hyper-dimensional imputation, and prioritize weight interpretation of features while still maintaining crucial information about a patient's health conditions and progress. The pharmacologic class of a drug refers to the grouping of medications with active ingredients that are similar based on a combination of any three attributes: their mechanism of action, their physiologic effect, and their chemical structure [23].

2.2. Model construction

The overall modeling and hyper-parameter tuning process is illustrated in Fig. 2B. We adopted two machine learning algorithms, multivariate Logistic Regression (LR) and Random Forest (RF), to build the models. The two algorithms were chosen to prioritize human interpretability [24], helping physicians identify features of interest in a straightforward manner, and potentially suggesting actionable items. In particular, the numerical values of Logistic Regression coefficients can indicate their magnitudes of influence over prediction outcomes [25], and the Gini indices of Random Forest are linked to how the "tree" evaluates splitting of its "branches" at each categorical predictor as it approaches final classification [26]. These relations can intuitively reveal important features and are more friendly to human comprehension compared to "black box" deep learning, whose model weights and biases do not elucidate their predictions [27]. Another related reason we chose the two algorithms over other time-series techniques such as recurrent neural networks [28] or transformers [29] is due to the lack of a suitable, interpretable preprocessing approach that could be applied across our patient sample. The data were collected from patients with past hospital admissions, whose both length of stays and number of stays varied. Such differences in data frequency would require extensive imputation work [30], and may further obfuscate any human attempt to comprehend the connection between the original data input and the final prediction output.

We also constructed an ensemble model (Ensemble) of LR and RF, to increase the discrimination capacity [31, 32]. That is, for each patient, we adopted the Average Ensemble [33], which averaged the two prediction scores yielded from the LR and RF models and used this new average as the patient's final predicted score for CDI positive test result. Aside from the potential gain in predictive performance, this might help confirm whether the raw prediction scores of LR and RF were generally complementary (Ensemble AUC would increase, or at least not be substantially worse than LR/RF results) or contradictory (Ensemble AUC would show significant drop).

To tune the hyper-parameters of the LR and RF models, we shuffled and held out 10 % of the dataset for testing purposes (Appendix Figure A.1A). It is noted that due to the high imbalance between positives and controls, the loss function used proportionally-assigned weights to prevent the models from overlooking the minority class. Tuning of hyper-parameters for both algorithms was done via grid search using the 10-fold cross-validation procedure on the remaining 90 % data; at each fold, we trained on 81 % and validated on 9 % of data (Appendix Figure A.1B). With the selected hyper-parameters (details in Appendix Table A.1), two models (LR and RF) were retrained on 90 % of data (excluding the hold-out test set), and the performance of our models are evaluated on the hold-out 10 % of data (Appendix Figure A.1C).

For implementation, we employed the Scikit-learn framework [34] for computations. We also adopted libraries for visualization (Matplotlib) [35] and statistical analysis (SpiCy) [36]. We used the Python programming language. The coding task was done exclusively within a secured, HIPAA-compliant virtual environment.

2.3. Evaluation

The overall evaluation process is shown in Fig. 2C. For model performance, we used the Area Under the Receiver Operating

Characteristic Curve (AUROC) [37] as our evaluation metric. We reported two sets of AUROCs for each of the three algorithms (LR, RF and Ensemble). (1) *Cross-Validation AUROC*. We averaged the AUROCs of each algorithm's respective 10 models that were constructed in the cross-validation step (that is, models that were built on 81 % of training data using the selected hyper-parameters, then tested on their respective folds of 9 % data). (2) *Test AUROC*. We recorded the AUROC of each algorithm's *final model* that was built on 90 % of data and tested on the 10 % hold-out set.

Additionally, we conducted exploratory data analysis to understand the demographic characteristics of the study population. Afterwards, on the LR final model, we performed coefficient significance analysis to record the magnitudes and corresponding odds ratios (which were mathematically derived from and thus representative of the coefficient magnitudes), and p-values. For this coefficient significance assessment, we first conducted a univariate analysis, then a multivariate analysis. The initial step of univariate analysis was simply to confirm the overlapping of magnitude and statistical significance of each predictor with the results seen in the main multivariate approach, if any, other than to exclude predictors with high *P-values* from being considered in the multivariate analysis step. In addition, as it has been shown that correlations among classes of drugs may contribute to certain changes in a patient's gut health [38, 39] and enhance toxin production in some strains [40], and thus may affect the chance of CDI in that patient, we did not test for covariate independence. On the RF model, we recorded the Gini importance index [41] of each feature, to capture variables that may hold more weights in affecting the risk of CDI [42]. We also conducted a decision boundary analysis to reevaluate the conventional discrimination threshold of 0.5, taking into consideration that physicians may want to be more proactive and assertive when it comes to weighing the cost of a false positive versus that of a false negative prediction. For the three final models, we used the shift in predictions as this threshold changes between 0 and 1 at intervals of 0.01, by recording the pairwise numbers of false positive and false negative cases predicted at each threshold. This was for easy visualization of such trade-offs as compared to the ROC (Receiver Operating Characteristic) curve.

3. Results

3.1. Demographic characteristics

The distributions of gender (female versus non-female, which may include unidentified genders) between the Positive and Control groups did not seem to have a stark difference. Meanwhile, the proportion of patients in the Positive group that were White was statistically higher than that of the Control group, and so was the mean age of the Positive group (58.43) versus the Control group (53.59). With regards to their medical histories, on average, patients in the Positive group had more medication units per patient. Table 1 summarizes the patient characteristics of our dataset.

3.2. Model performance

For the *Cross-Validation AUROC*, the LR algorithm yielded an average AUROC of 0.793 (95 % Confidence Interval (CI) = (0.763, 0.823)); the RF algorithm's average AUROC was 0.833 (95 % CI = (0.805, 0.861)); and the Ensemble' average AUROC was 0.828 (95 % CI = (0.802, 0.854)). For the *Test AUROC* at the step of testing on the hold-out set, the final LR and RF models achieved AUROCs of 0.839 and 0.851, respectively. The final Ensemble model that combined the decisions of LR and RF algorithms yielded an AUROC of 0.866 (Fig. 3). In short, all our three constructed models demonstrated strong predictive powers, as shown by their AUROC scores along either construction step.

3.3. Feature analysis

For feature analysis, Table 2 shows a list of top 20 features as ranked by the lowest p-values ($p < 0.0001$), and their corresponding LR odds ratios in the multivariate analysis step (Table 2A). Similarly, Table 2B shows 20 features as ranked by the highest Gini indices. See Appendix B for the complete list of features, which entails their corresponding AUROCs, p-values and odds ratios in the LR univariate analysis step; their p-values and odds ratios in the LR multivariate analysis step; and the ranking of their corresponding Gini indices.

Table 1
Demographic backgrounds of the two Positive and Control groups.

Patient Characteristics	Positive	Control
Number of cases (N)	1541	155,952
Gender female vs non-female (n)	734 (47.63 %)	77,463 (49.67 %)
Race White vs non-White (n)*	923 (59.9 %)	87,130 (55.9 %)
Mean age (mean, standard deviation)*	58.43 (17.23)	53.59 (18.99)
Median age	60.2	54.93
Number of medication units (mean, standard deviation)*	84.95 (107.3)	28.77 (57.84)

Asterisks ("*") indicate Statistically different at $p < 0.001$.

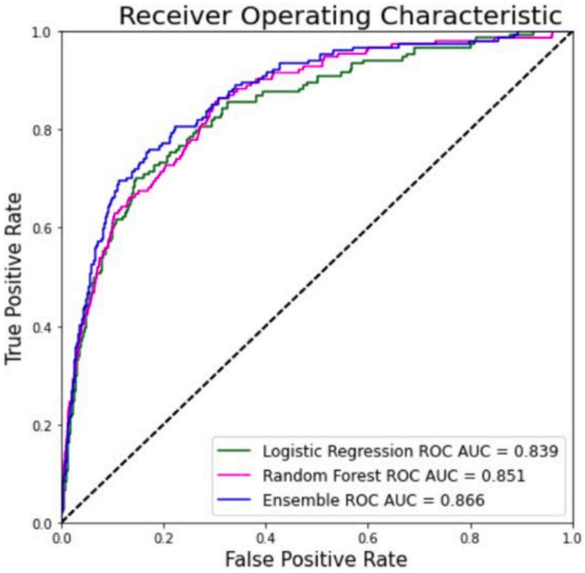

Fig. 3. AUROC curves of three final models: Logistic Regression, Random Forest, and Ensemble.

Table 2
Results of feature analysis on the Logistic Regression and Random Forest models. A) Top 20 features with $p < 0.001$ in LR multivariate analysis ranked by the lowest *P-values*. B) Top 20 features as ranked by their RF Gini indices.

A) Logistic Regression Multivariable Analysis		B) Random Forest Feature Importance	
Feature	Odds Ratio	Feature	Gini Index
Age	1.0144	Minerals & electrolytes	0.1507
Misc. anti-infectives	1.2558	Misc. anti-infectives	0.1493
Ophthalmic	0.8889	Unassigned group	0.0686
Anticoagulants	0.9377	Antiemetics	0.0396
Misc. GI	1.1577	Analgesics - opioids	0.0369
Fluoroquinolones	1.2250	Diuretics	0.0319
Tetracyclines	0.7735	Age	0.0295
Analgesics - opioids	1.0242	Anticoagulants	0.0261
Local anesthetics - parenteral	1.0816	Local anesthetics - parenteral	0.0256
Minerals & electrolytes	1.0265	Fluoroquinolones	0.0190
Toxoids	0.5740	Assorted classes	0.0190
Laxatives	0.9268	Misc. GI	0.0179
Antidiarrheals	2.3529	Antihistamines	0.0172
Anti-rheumatic	0.9299	Penicillins	0.0171
Gout	1.2108	Corticosteroids	0.0164
Antineoplastics	0.9684	Ulcer drugs	0.0158
Unassigned group	1.0391	Hematopoietic agents	0.0149
Penicillins	1.1491	Misc. Hematological	0.0147
Diagnostic products	0.9326	Antineoplastics	0.0144
Other or mixed races	0.8294	Analgesics - non-opioids	0.0142

3.4. Decision boundary analysis

The trade-off curves (Fig. 4) show the effects as the threshold changes between 0.4 and 0.6 at intervals of 0.01, for each of the models LR (Fig. 4A), RF (Fig. 4B) and Ensemble (Fig. 4C). Each point on the curve illustrates the paired numbers of false negatives and false positives that the Ensemble model would have predicted on the hold-out set at that threshold. See Appendix C for the full trade-off curve with decision boundaries ranging from 0 to 1.

4. Discussion

4.1. Findings

In general, the high AUROC scores of three models that persisted in both the cross-validation and the testing phase imply that they can be adopted to support the prediction of positive CDI test results. During cross-validation, the average AUROC of the RF algorithm

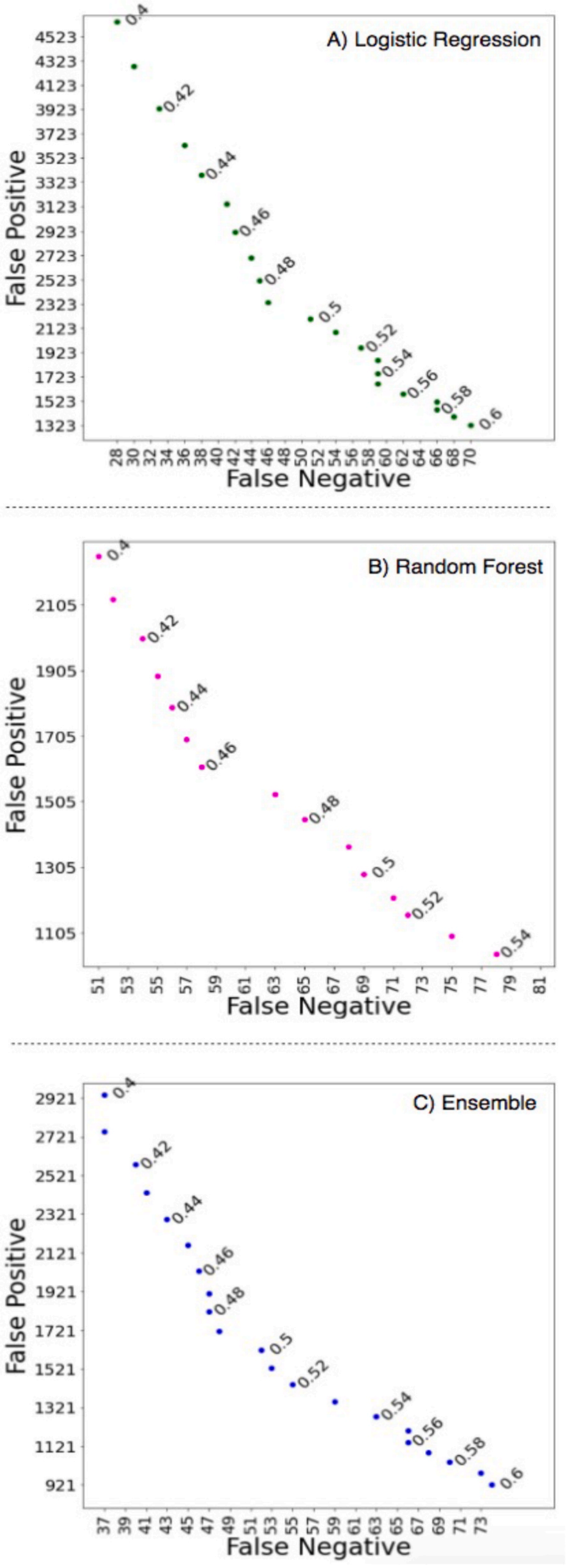

Fig. 4. The trade-off between the number of false negatives and false positive cases as the discrimination threshold changes between 0.4 and 0.6 for each algorithm. **(A)** Logistic Regression. **(B)** Random Forest. **(C)** Ensemble.

was highest among the three models, while during testing the Ensemble algorithm outperformed the other two. In general, they all show relatively high discrimination results (AUROC between 0.793 and 0.866), and the fact that Ensemble AUC did not exhibit a stark drop shows that the two LR and RF algorithms tended to yield complimentary predictions. These AUROC scores are higher than those obtained from models built on smaller sample sizes and/or with more covariates (0.75–0.82 of AUROC in existing studies [18 20]).

Among covariates with a low p-value in the LR final model, there are clinically seen factors that may affect a patient's CDI risk such as age [43], historic use of drugs in the anti-infective class (penicillins, fluoroquinolones) [44–46], implied immune-compromise conditions through cancer medications and their subsequent side effects of diarrhea resulting in more testing [47], and indications of past gastrointestinal (GI) issues (Misc. GI). This agreement is again repeated in the RF model, shown through the relatively similar ranking of their Gini indices. The consistency with regards to which feature may have a higher predictive power among such approaches suggests that the models are clinically useful for physicians. Moreover, the high impact of antibiotic use is again confirmed, which may help to caution physicians against over-prescribing of antibiotics for susceptible patients. A study finds that reducing patient's susceptibility to CDI, with which antibiotic treatments can increase risk of colonization and subsequent progression to disease [48], is more effective in combating CDI as compared to lowering the transmission rate [49].

The discrimination threshold analysis curves, on the other hand, render a visually intuitive tool to aid physicians, infection control professionals and laboratory medicine teams in deciding their own testing policies. Specifically, these team members can use the curves to precisely determine an appropriate threshold value that can minimize the number of false negatives while reducing false positives as compared to their neighbors, such as the threshold of 0.55 for Logistic Regression (about 100 false positive cases reduced), or 0.48 and 0.56 for the Ensemble model (about 50–100 false positives reduced). Hospital administrators can also use the curves in their cost analysis studies [50]; for example, the threshold values can help estimate the savings or expenses associated with reassessment of diagnosis [51]. It is known that in terms of healthcare cost, a CDI-positive inpatient with health plan incurs about $21,000 more in treatment charge than an inpatient without CDI [52]. This number is even higher when the CDI is recurrent [53]. At the same time, the cost of a CDI stool test ranges from $15 to $128 (pricing current as of 2021) [54]. Therefore, when coupled with experiences and real-life knowledge of previous cases in their practices, a threshold-adjusted model may supply healthcare providers with discretionary insights to balance between over-treatment and under-diagnosing. This is a lesson transferable to other serious infections where any false negative case may cause great, unintended harm, yet the effort to minimize false negative might trigger more false positives than desired, such as COVID-19 [55].

From a practical viewpoint, the predictions may help guide clinical workflow decisions, such as the utilization of preemptive decontamination methods on both organic and inanimate surfaces, and better tracing of patient movement within the care facility, should a patient be considered of higher risk. For example, a recent study traced CDI back to a single CT scanner in a large academic hospital [56], highlighting the need for preventative measures even without obvious signs of spread.

4.2. *Limitations*

The limitations of our study include:

(1) *Selection of algorithms and model calibration.* To encourage ease of intuitive human interpretability, we used LR, RF and Ensemble which are well-known highly interpretable predictive models. We are yet to utilize other ML algorithms such as deep learning (e.g., RNNs and transformers), nor to adopt other ensemble methods such as boosting [57] and stacking [58]. We are also yet to measure the calibration of our models using metrics such as estimated calibration index (ECI) [59], to calibrate our models using methods such as isotonic regression [60, 61], or to experiment on different data balancing strategies such as downsampling/upsampling [62].

(2) *Potential feature codependency.* We did not investigate possible intercorrelations among pharmacologic class predictors in our models. This may need further analysis to see how specific interactions affect CDI risk.

(3) *In-practice deployment.* We did use real patient data to build our models, while deploying our models in a true clinical workflow requires more investigation. Further work may be needed to validate whether our predictions would apply to high-risk patients who would conventionally not be tested by the clinical teams.

(4) *Deployment and broader evaluation of the predictive models.* Currently, we only evaluated our model on retrospective patient data from UCSD. We are yet to deploy the model at UCSD or evaluate the performance of the models in other institutions or countries.

(5) *Distinction between colonization and infection.* The models predict positive test results, both true infection and colonization with *C. diff.* While colonization does have a strong positive association with true infection, this relationship is not perfect. We are yet to further analyze the correlation between colonization and explicit infection.

(6) *Cost analysis.* We have yet to investigate the potential cost associated with the adjustment of the discrimination threshold, such as information on the labor and financial expenses associated with surveilling near-threshold patients, and to yield a full cost analysis.

5. Conclusions

Our machine learning models using large-scale, longitudinal patients' medication history and demographic backgrounds can predict the first positive CDI test results with high performance. Our models are built on a large dataset of 157493 real-world UCSD Health patients and capture the underlying demographic and medical information of each patient's health over a 3-year time window, using a reasonable number of 104 covariates. The models achieve strong AUROCs between 0.839 and 0.866, and their face validity is highlighted as they find clinically-attested factors that may increase a patient's risk for CDI, such as age, antibiotic uses, cancer and/or GI-related conditions. Their discrimination threshold analysis aids physicians in deciding their own level of precaution. This is a transferable lesson to other infectious diseases such as Coronavirus Disease 2019 (COVID-19), as it considers the starkly different weights of false negatives and false positives. Specifically, as both CDI and COVID-19 tend to affect people with pre-existing conditions and/or comorbidity, early screening using our developed predictive models may help prevent disease outbreaks and mitigate severe consequences beyond the initial infections. This in turn could improve patient care, limit healthcare-associated cost, and boost hospital performance with regards to meeting CDC goals.

CRediT authorship contribution statement

Anh Pham: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Robert El-Kareh:** Writing – review & editing, Conceptualization. **Frank Myers:** Writing – review & editing. **Lucila Ohno-Machado:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Tsung-Ting Kuo:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Conceptualization.

Ethics and consent

The Human Research Protection Program at the University of California, San Diego approved this project and granted it a waiver of informed consent on 05/17/2019 (IRB Project #190457CX).

Funding

The authors were funded by the US National Institutes of Health (NIH) [R01EB031030] and the Graduate Division San Diego Matching Fellowship associated with San Diego Biomedical Informatics Education & Research (SABER) NIH National Library of Medicine (NLM) [grant T15LM011271]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Michael Hogarth, MD and UCSD ACTRI for the technical support for the VRD computing environment, and Paulina Paul, MS for the help to extract Epic data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e41350.

References

[1] Fordtran JS. Colitis Due to Clostridium Difficile Toxins: Underdiagnosed, Highly Virulent, and Nosocomial. https://doi.org/10.1080/08998280.2006.119281142017 doi: 10.1080/08998280.2006.11928114.

[2] R. M, W. Mh, G. Dn, Clostridium difficile infection: new developments in epidemiology and pathogenesis, Nat. Rev. Microbiol. 7 (7) (2009), https://doi.org/10.1038/nrmicro2164.

[3] R.L. Jump, C.J. Donskey, Clostridium difficile in the long-term care facility: prevention and management, Current geriatrics reports 4 (2015) 60–69.

[4] Clostridioides difficile Infection. Secondary Clostridioides difficile Infection 2020-01-02T08:17:57Z 2020. https://www.cdc.gov/hai/organisms/cdiff/cdiff_infect.html.

[5] Prevention CDC, Antibiotic resistance threats in the United States, 2019. Secondary Antibiotic Resistance Threats in the United States, 2019 2019.

[6] Current HAI Progress Report, Secondary Current HAI Progress Report 2021-02-22T01:37:08Z, 2019. https://www.cdc.gov/hai/data/portal/progress-report.html.

[7] M. Alrawashdeh, C. Rhee, H. Hsu, R. Wang, K. Horan, G.M. Lee, Assessment of federal value-based incentive programs and in-hospital Clostridioides difficile infection rates, JAMA Netw. Open 4 (10) (2021) e2132114.

[8] Healthcare-Associated Infections Report UC Sand Diego Health. Secondary Healthcare-Associated Infections Report UC San Diego Health, 2017.

[9] D. C, H. Lp, B. R, J. Lt, Biocide resistance and transmission of Clostridium difficile spores spiked onto clinical surfaces from an American health care facility, Appl. Environ. Microbiol. 85 (17) (2019), https://doi.org/10.1128/AEM.01090-19.

[10] A.N. Edwards, S.T. Karim, R.A. Pascual, L.M. Jowhar, S.E. Anderson, S.M. McBride, Chemical and stress resistances of Clostridium difficile spores and vegetative cells, Front. Microbiol. 7 (2016) 1698.

[11] A. Rineh, M.J. Kelso, F. Vatansever, G.P. Tegos, M.R. Hamblin, Clostridium difficile infection: molecular pathogenesis and novel therapeutics, Expert Rev. Anti-infect. Ther. 12 (1) (2014) 131–150.

[12] Y. Longtin, B. Paquet-Bolduc, R. Gilca, et al., Effect of detecting and isolating Clostridium difficile carriers at hospital admission on the incidence of C difficile infections: a quasi-experimental controlled study, JAMA Intern. Med. 176 (6) (2016) 796–804, https://doi.org/10.1001/jamainternmed.2016.0177.
[13] H.S. Lee, K. Plechot, S. Gohil, J. Le, Clostridium difficile: diagnosis and the consequence of over diagnosis, Infectious diseases and therapy (2021) 1–11.
[14] Clostridium difficile testing guidance. In: Health WSDo, ed.
[15] I.M. Zacharioudakis, F.N. Zervou, E.E. Pliakos, P.D. Ziakas, E. Mylonakis, Colonization with toxinogenic C. difficile upon hospital admission, and risk of infection: a systematic review and meta-analysis, Official journal of the American College of Gastroenterology| ACG 110 (3) (2015) 381–390.
[16] S.R. Curry, C.A. Muto, J.L. Schlackman, et al., Use of multilocus variable number of tandem repeats analysis genotyping to determine the role of asymptomatic carriers in Clostridium difficile transmission, Clin. Infect. Dis. 57 (8) (2013) 1094–1102.
[17] J.S. Biswas, A. Patel, J.A. Otter, E. van Kleef, S.D. Goldenberg, Contamination of the hospital environment from potential Clostridium difficile excretors without active infection, Infect. Control Hosp. Epidemiol. 36 (8) (2015) 975–977.
[18] J. Wiens, E. Horvitz, J. Guttag, Patient risk stratification for hospital-associated C. Diff as a time-series classification task, Adv. Neural Inf. Process. Syst. 25 (2021).
[19] E.R. Dubberke, Y. Yan, K.A. Reske, et al., Development and validation of a Clostridium difficile infection risk prediction model, Infect. Control Hosp. Epidemiol. 32 (4) (2011) 360–366.
[20] J. Oh, M. Makar, C. Fusco, 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. 39 (4) (2018) 425–433.
[21] D.A. Katz, M.E. Lynch, B. Littenberg, Clinical prediction rules to optimize cytotoxin testing for Clostridium difficile in hospitalized patients with diarrhea, Am. J. Med. 100 (5) (1996) 487–495.
[22] Early Prediction of Positive Clostridioides Difficile Test Results, AMIA Annual Symposium, 2021.
[23] Pharmacologic Class. Secondary Pharmacologic Class. https://www.fda.gov/industry/structured-product-labeling-resources/pharmacologic-class.
[24] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, L. Cilar, Interpretability of machine learning-based prediction models in healthcare, Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 10 (5) (2020) e1379.
[25] S. Sperandei, Understanding logistic regression analysis, Biochem. Med. 24 (1) (2014) 12–18.
[26] S. Nembrini, I.R. König, M.N. Wright, The revival of the Gini importance? Bioinformatics 34 (21) (2018) 3711–3718.
[27] D. Castelvecchi, Can we open the black box of AI? Nature News 538 (7623) (2016) 20.
[28] L.R. Medsker, L. Jain, Recurrent neural networks, Design and Applications 5 (64–67) (2001) 2.
[29] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).
[30] P.B. Weerakody, K.W. Wong, G. Wang, W. Ela, A review of irregular time series data handling with gated recurrent neural networks, Neurocomputing 441 (2021) 161–178.
[31] Ensembles of nlp tools for data element extraction from clinical notes, AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2016.
[32] T.-T. Kuo, J. Kim, R.A. Gabriel, Privacy-preserving model learning on a blockchain network-of-networks, J. Am. Med. Inf. Assoc. 27 (3) (2020) 343–354, https://doi.org/10.1093/jamia/ocz214.
[33] M.M. Li, A. Pham, T.-T. Kuo, Predicting COVID-19 county-level case number trend by combining demographic characteristics and social distancing policies, JAMIA open 5 (3) (2022) ooac056.
[34] Scikit-learn, Machine learning in Python. Secondary scikit-learn: machine learning in Python. https://scikit-learn.org/stable/, 2022.
[35] Matplotlib: visualization with Python. Secondary matplotlib: visualization with Python. https://matplotlib.org/, 2022.
[36] SciPy, 2022.
[37] T.A. Lasko, J.G. Bhagwat, K.H. Zou, L. Ohno-Machado, The use of receiver operating characteristic curves in biomedical informatics, J. Biomed. Inf. 38 (5) (2005) 404–415.
[38] M. Klünemann, S. Andrejev, S. Blasche, et al., Bioaccumulation of therapeutic drugs by human gut bacteria, Nature 597 (7877) (2021) 533–538, https://doi.org/10.1038/s41586-021-03891-8.
[39] F. Imhann, A. Vich Vila, M.J. Bonder, et al., The influence of proton pump inhibitors and other commonly used medication on the gut microbiota, Gut Microb. 8 (4) (2017) 351–358.
[40] M.J. Aldape, A.E. Packham, D.W. Nute, A.E. Bryant, D.L. Stevens, Effects of ciprofloxacin on the expression and production of exotoxins by Clostridium difficile, J. Med. Microbiol. 62 (Pt 5) (2013) 741.
[41] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Adv. Neural Inf. Process. Syst. 26 (2013) 431–439.
[42] S.L. Baxter, C. Marks, T.-T. Kuo, L. Ohno-Machado, R.N. Weinreb, Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records, Am. J. Ophthalmol. 208 (2019) 30–40.
[43] J. Pépin, L. Valiquette, B. Cossette, Mortality attributable to nosocomial Clostridium difficile–associated disease during an epidemic caused by a hypervirulent strain in Quebec, CMAJ (Can. Med. Assoc. J.) 173 (9) (2005) 1037–1042.
[44] A. Deshpande, V. Pasupuleti, P. Thota, et al., Community-associated Clostridium difficile infection and antibiotics: a meta-analysis, J. Antimicrob. Chemother. 68 (9) (2013) 1951–1961.
[45] J. Pépin, N. Saheb, M.-A. Coulombe, et al., Emergence of Fluoroquinolones as the predominant risk factor for Clostridium difficile–associated diarrhea: a cohort study during an epidemic in quebec, Clin. Infect. Dis. 41 (9) (2005) 1254–1260, https://doi.org/10.1086/496986.
[46] K.Z. Vardakas, K.K. Trigkidis, E. Boukouvala, M.E. Falagas, Clostridium difficile infection following systemic antibiotic administration in randomised controlled trials: a systematic review and meta-analysis, Int. J. Antimicrob. Agents 48 (1) (2016) 1–10.
[47] D.A. Leffler, J.T. Lamont, Clostridium difficile infection, N. Engl. J. Med. 372 (16) (2015) 1539–1548.
[48] P. Warn, P. Thommes, A. Sattar, et al., Disease progression and resolution in rodent models of Clostridium difficile infection and impact of antitoxin antibodies and vancomycin, Antimicrob. Agents Chemother. 60 (11) (2016) 6471–6482.
[49] J. Starr, A. Campbell, E. Renshaw, I. Poxton, G. Gibson, Spatio-temporal stochastic modelling of Clostridium difficile, J. Hosp. Infect. 71 (1) (2009) 49–56.
[50] Jr LP. Garrison, J.B. Babigumira, A. Masaquel, B.C. Wang, D. Lalla, M. Brammer, The lifetime economic burden of inaccurate HER2 testing: estimating the costs of false-positive and false-negative HER2 test results in US patients with early-stage breast cancer, Value Health 18 (4) (2015) 541–546.
[51] J.E. Lafata, J. Simpkins, L. Lamerato, L. Poisson, G. Divine, C.C. Johnson, The economic impact of false-positive cancer screens, Cancer Epidemiol. Biomark. Prev. 13 (12) (2004) 2126–2132.
[52] S. Zhang, S. Palazuelos-Munoz, E.M. Balsells, H. Nair, A. Chit, M.H. Kyaw, Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study, BMC Infect. Dis. 16 (1) (2016) 1–18.
[53] R. Rodrigues, G.E. Barber, A.N. Ananthakrishnan, A comprehensive study of costs associated with recurrent Clostridium difficile infection, Infect. Control Hosp. Epidemiol. 38 (2) (2017) 196–202.
[54] How Much Does an Stool, C-Diff Test Cost Near Me? - MDsave, Secondary How Much Does an Stool, C-Diff Test Cost Near Me? - MDsave (2021), in: https://www.mdsave.com/procedures/stool-c-diff-test/d787ffc4.
[55] T. Dai, S. Singh, Overdiagnosis and undertesting for infectious diseases, Market. Sci. (2024), https://doi.org/10.1287/mksc.2022.0038.
[56] S.G. Murray, J.W.L. Yim, R. Croci, et al., Using spatial and temporal mapping to identify nosocomial disease transmission of Clostridium difficile, JAMA Intern. Med. 177 (12) (2017) 1863–1865, https://doi.org/10.1001/jamainternmed.2017.5506.
[57] A. Mayr, H. Binder, O. Gefeller, M. Schmid, The evolution of boosting algorithms, Methods Inf. Med. 53 (6) (2014) 419–427.
[58] S. Džeroski, B. Ženko, Is combining classifiers with stacking better than selecting the best one? Mach. Learn. 54 (2004) 255–273.
[59] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cack, M.J. Pencina, E.W. Steyerberg, A calibration hierarchy for risk models was defined: from utopia to empirical data, J. Clin. Epidemiol. 74 (2016) 167–176.

[60] Reliably calibrated isotonic regression, Pacific-asia Conference on Knowledge Discovery and Data Mining, Springer, 2021.

[61] M. Edelson, T.-T. Kuo, Generalizable prediction of COVID-19 mortality on worldwide patient data, JAMIA open 5 (2) (2022) ooac036.

[62] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem. Int. J. Advance Soft Compu, Appl 5 (3) (2013) 176–204.