Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2022, Volume 2, Issue 1, Page No: 85-96

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Interpretable Machine Learning Prediction of Clostridioides difficile Infection Using Three-Year Longitudinal EHR Data

John Peterson[1], Mark Reynolds[1], Kevin Brooks[1]*

[1]Department of Health Systems Science, School of Medicine, University of Michigan, Ann Arbor, USA.

***E-mail** ✉ kevin.brooks.hs@gmail.com

Received: 08 January 2022; **Revised:** 11 March 2022; **Accepted:** 11 Marh 2022

ABSTRACT

Clostridioides difficile infection poses major clinical and operational challenges. Hospitals have both quality and economic motivations to manage CDI effectively. Universal admission screening is rarely recommended, and prior modeling efforts often relied on limited samples, overly complex feature sets, or black-box techniques. Our goal was to create models using patient information to estimate the likelihood of a positive test with strong discrimination, clear interpretability, and a practical set of long-term health indicators. We used records from 157,493 UC San Diego Health patients seen between January 01, 2016, and July 03, 2019 who had at least 6 months of medication history. Pregnant individuals, patients under 18, and incarcerated persons were excluded. We trained Logistic Regression, Random Forest, and Ensemble models using hyperparameters tuned through 10-fold cross-validation. Performance was evaluated by AUROC. Logistic Regression coefficients were examined via odds ratios and p-values; Random Forest feature contributions were assessed using Gini importance. We also compared false-positive and false-negative predictions at selected thresholds.

The Logistic Regression, Random Forest, and Ensemble models produced AUROCs of 0.839, 0.851, and 0.866, respectively. Variables associated with elevated risk included age, use of immunosuppressive therapies, previous antibiotic exposure, and certain gastrointestinal medications. All models demonstrated strong discrimination (AUROC >0.83). Across analytic methods, similar predictors emerged as influential, many of which are consistent with established clinical risk factors for Clostridioides difficile. These human-readable models help identify factors shaping a patient's likelihood of a positive test and the associated infection risk.

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

How to Cite This Article: Peterson J, Reynolds M, Brooks K. Interpretable Machine Learning Prediction of Clostridioides difficile Infection Using Three-Year Longitudinal EHR Data. Interdiscip Res Med Sci Spec. 2022;2(1):85-96. https://doi.org/10.51847/do0gNijk3T

Introduction

Clostridioides difficile infection and diagnostic complexity

CDI, caused by C. diff, can lead to severe gastrointestinal disease, including colitis, pseudomembranous colitis, life-threatening diarrhea, and sepsis [1, 2]. Older adults and those exposed to antibiotics—particularly in long-term care—are at amplified risk [3]. The CDC classifies CDI as a significant national threat [4]; in 2017, U.S. hospitals recorded roughly 223,900 cases, 12,800 deaths, and close to $1 billion in HA-related costs [5]. Not meeting the CDC's standardized infection ratio (SIR) [6] can harm a hospital's standing and impose financial burdens [7]. For instance, UCSD Health's 2015–2017 HA-CDI rate exceeded the 2015 national baseline SIR [8]. As a result, lowering HA-CDI is a central quality and financial goal.

C. diff spores withstand many common disinfectants and persist on treated surfaces [9–11], motivating interest in more proactive strategies that remain aligned with guidelines. One institution reported that screening nearly all admissions prevented up to 62 % of expected infections and steadily reduced CDI rates [12]. Early case identification also clarifies whether infections are hospital- or community-acquired, improving monitoring accuracy. Nonetheless, broad admission testing is discouraged because it may lead to overdiagnosis and unneeded

antimicrobial therapy, accelerating resistance [13]. Current recommendations, therefore, emphasize testing only when symptoms appear [14]. This creates a timing dilemma: patients colonized at admission have higher odds of developing CDI [15] and can shed spores that contribute to transmission [16, 17]. These constraints highlight the value of data-driven methods capable of estimating risk without requiring immediate laboratory testing, while still offering actionable insights to clinicians.

Current machine learning models to predict CDI and their drawbacks

Researchers have proposed various risk-stratification tools to help identify individuals likely to test positive for C. diff. However, most prior efforts relied on datasets with relatively small patient cohorts (roughly 8,000–36,000 individuals [18, 19]); very large feature sets (around 1,800–5,000 variables—often binary encodings of attributes that only apply to some patients [20]), which make the resulting models difficult to interpret; and/or predictors tied heavily to recent clinical activity, such as antibiotic prescriptions within 30 days of testing [21], which may overlook the cumulative impact of microbiome-altering treatments over longer intervals [22].

To address these shortcomings, we built CDI prediction models that (i) use a substantially larger real-world cohort (157,493 UCSD Health patients), (ii) offer strong discrimination and interpretability by limiting the model to 104 core demographic and medication-based predictors, and (iii) capture longitudinal health history over a 3-year period. Because a positive laboratory result may reflect colonization rather than active CDI, our intention is to provide clinicians with an additional decision-support tool—one that complements existing diagnostic pathways and helps balance the gap between testing every asymptomatic patient and waiting for severe symptoms before screening.

Objective

Our task is to predict the first instance of a positive C. diff laboratory test using demographic variables and medication history up to one calendar day before the order date for that first positive test **(Figure 1)**. In practice, this typically falls at least two days before results are available, potentially enabling earlier intervention. Since adults carrying toxigenic C. diff have a six-fold greater risk of progressing to infection [15], anticipating colonization may play an important role in reducing CDI incidence.

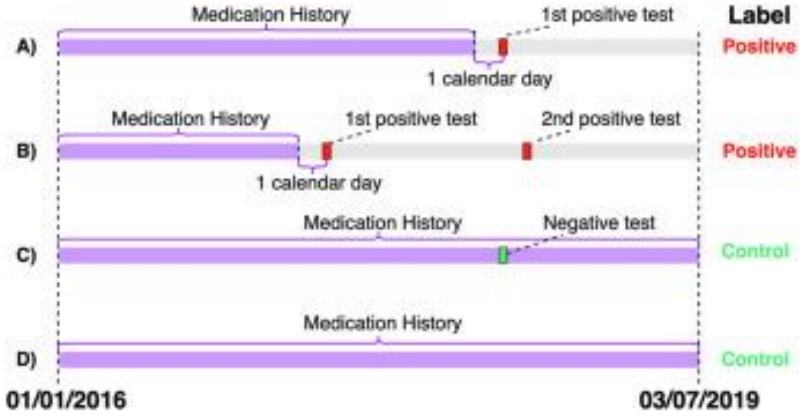


Figure 1. Overview of the prediction labels used in this study. Four scenarios define patient classification: (A) a patient with a single positive test is categorized as Positive; (B) a patient with several positive tests is also labeled Positive, using the earliest result; (C) a patient with a negative test is classified as a Control; and (D) a patient who never receives a C. diff test is also treated as a Control.

Materials and Methods

The study pipeline is summarized in **Figure 2**, consisting of three major steps: Data, Model Construction, and Evaluation. Each component is explained in the subsections below.

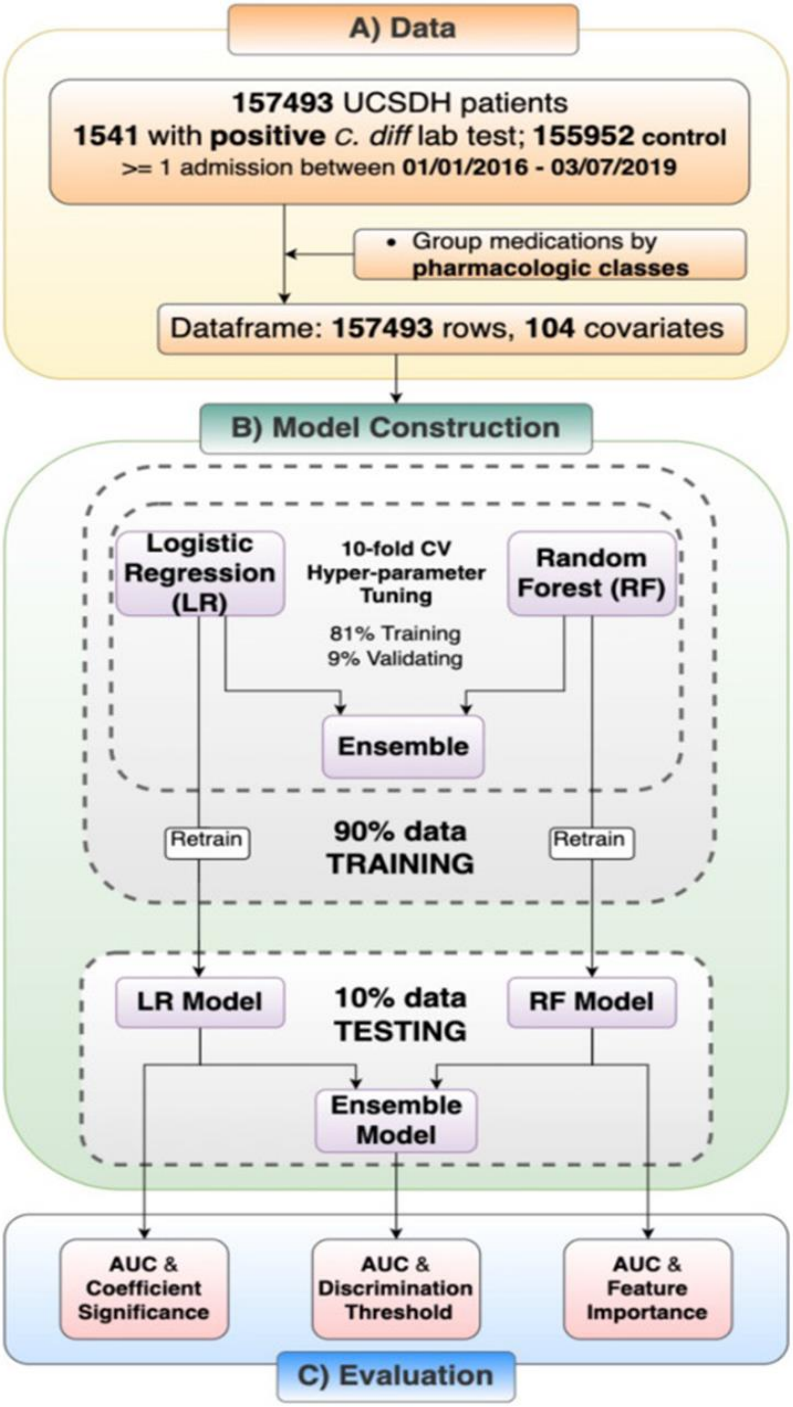


Figure 2. Study workflow.
(A) Data: Demographics and medication histories for 157,493 UCSD Health patients seen between 01/01/2016–03/07/2019 were transformed into analytic vectors.
(B) Model Construction: We developed models using Logistic Regression, Random Forest, and an Ensemble approach.
(C) Evaluation: We assessed model performance and conducted feature-level analyses.

Data

Figure 2a outlines the data preparation process. We accessed information for 157,493 UCSD Health patients admitted between January 01, 2016 and July 03, 2019 (IRB approval #190457CX). Eligible patients were those with at least one admission during the study period and a minimum of six months of medication history. Excluded groups included individuals younger than 18, pregnant patients, incarcerated persons, and those missing age data. In total, 1,541 individuals (1%) were labeled as having a CDI Positive test result as defined in **Figure 1**.

To construct the predictors, we selected a focused set of variables covering demographic factors and medication exposures—attributes that naturally reflect a patient's physiologic trajectory over time. The final dataset contained 104 covariates: 10 demographic variables (age, race, gender, etc.) derived from UCSD Health metadata, and 94 medication-related features summarizing usage patterns during the observation window. Medication names (originally 4,420 unique items) were consolidated into 94 pharmacologic classes to reduce dimensionality, avoid extensive imputation, improve interpretability, and retain essential clinical meaning. A pharmacologic class groups medications whose active components share similarities based on one or more characteristics: mechanism of action, physiologic impact, or chemical structure [23].

Model construction

The model-building and hyperparameter-selection workflow is depicted in **Figure 2b**. We selected two algorithms—multivariate Logistic Regression (LR) and Random Forest (RF)—because they offer clearer interpretability for clinical users [24]. Their outputs allow clinicians to pinpoint influential factors and consider potential clinical actions. For LR, the magnitude and sign of each coefficient reflect its impact on predicted risk [25], while for RF, the Gini importance reflects how each variable contributes to the tree-based split decisions that lead to classification [26]. Both approaches offer more transparent logic than deep learning models, whose internal parameters typically provide little insight into how predictions are formed [27]. Another motivation for using LR and RF instead of temporal models such as recurrent neural networks [28] or transformers [29] is the absence of a coherent, interpretable preprocessing strategy for highly irregular patient histories. Patients differ widely in length of stay and number of admissions, and applying time-series models would require extensive imputation [30], further obscuring how original inputs relate to final predictions.

To further boost discrimination, we created an Ensemble method combining LR and RF [31, 32]. Specifically, we used an Average Ensemble [33], taking the mean of the LR and RF prediction scores to form a final risk estimate for each patient. Besides possibly improving accuracy, this also reveals whether the two algorithms provide complementary information (higher or stable AUC) or conflicting signals (notable drop in AUC).

To determine model hyperparameters, we randomly reserved 10% of the data as a held-out test set. Given the strong class imbalance, the loss function applied proportional class weights so the minority class was not ignored. Both LR and RF hyperparameters were tuned via grid search with 10-fold cross-validation on the remaining 90% of the dataset. During each fold, 81% of the data was used for training and 9% for validation. After identifying optimal settings, the LR and RF models were retrained using 90% of the full dataset, and performance was later evaluated on the 10% hold-out portion.

We implemented all models using Scikit-learn [34], produced visualizations with Matplotlib [35], and performed statistical computations with SpiCy [36]. All analyses were carried out in Python within a secure, HIPAA-compliant environment.

Evaluation

The evaluation procedure is summarized in **Figure 2c**. We used the Area Under the Receiver Operating Characteristic Curve (AUROC) [37] as the principal metric. For each algorithm (LR, RF, Ensemble), two AUROC values were reported:

1. Cross-Validation AUROC: The average AUROC across the 10 cross-validation models was trained on 81% of the data and evaluated on their respective 9% validation folds.
2. Test AUROC: The AUROC was calculated from the final model trained on 90% of the data and tested on the 10% hold-out set.

We also explored population demographics to contextualize the dataset. For the final LR model, we examined coefficient significance by comparing the estimated magnitudes, their associated odds ratios, and corresponding p-values. This included a univariate step followed by a multivariate analysis. The univariate stage served mainly to confirm consistency with multivariate findings and to avoid carrying forward predictors with very high p-values. Because interactions among drug classes may alter gut microbiota [38, 39] or enhance toxin production in certain strains [40], and thereby influence CDI risk, we did not enforce covariate independence.

For the RF model, we documented Gini importance scores [41] to identify features with stronger contributions to CDI risk [42]. Additionally, we performed a decision-boundary assessment to revisit the typical 0.5 classification threshold. Since clinicians may weigh false positives and false negatives differently, we examined how predictions change when the threshold moves from 0 to 1 in 0.01 increments. For each threshold, we recorded pairwise counts

of false positives and false negatives to allow clearer visualization of trade-offs than is available from the standard ROC curve.

Results and Discussion

Demographic characteristics

When comparing the Positive and Control cohorts, the overall split between female and non-female participants (including those without reported gender) showed no major contrast. In contrast, the Positive group contained a significantly larger fraction of White individuals, and their mean age was also higher (58.43 compared with 53.59). Regarding prior clinical records, individuals classified as Positive typically had a greater count of medication entries per person. A consolidated overview of these characteristics is presented in **Table 1**.

Table 1. Demographic overview for the Positive and Control cohorts.

Patient Characteristic	Positive C. difficile Cases (n = 1,541)	Control Group (n = 155,952)
Gender – Female	734 (47.63 %)	77,463 (49.67 %)
Race – White (vs non-White)*	923 (59.9 %)	87,130 (55.9 %)
Age (years) – Mean ± SD*	58.43 ± 17.23	53.59 ± 18.99
Age (years) – Median	60.2	54.93
Total medication units prescribed – Mean ± SD*	84.95 ± 107.3	28.77 ± 57.84

Asterisks ("∗") denote $p < 0.001$.

Model performance

During cross-validation, the Logistic Regression (LR) model achieved a mean AUROC of 0.793 (95% CI: 0.763–0.823). The Random Forest (RF) model reached 0.833 (95% CI: 0.805–0.861), while the Ensemble configuration obtained 0.828 (95% CI: 0.802–0.854). For the final evaluation on the untouched test partition, AUROC values were 0.839 for LR and 0.851 for RF. The combined Ensemble model—integrating outputs from LR and RF—yielded the highest AUROC at 0.866 **(Figure 3)**. Across both phases, each approach demonstrated strong discriminatory performance.

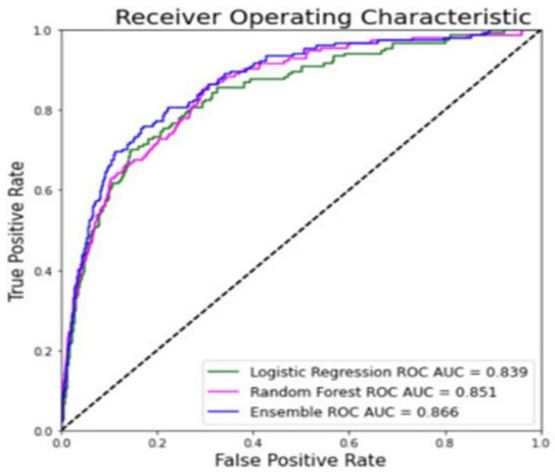


Figure 3. AUROC plots for the finalized LR, RF, and Ensemble models.

Feature analysis

Table 2a lists the twenty LR features showing the smallest p-values (all $p < 0.0001$) along with their multivariate odds ratios. **Table 2b** presents the twenty RF features with the highest Gini importance scores.

Table 2. Feature analysis results for LR and RF models. A) Top 20 LR multivariate predictors with $p < 0.001$, ordered by increasing p-value. B) Top 20 RF predictors ordered by Gini importance.

a) Logistic Regression – Multivariable Adjusted Odds Ratios		
Rank	**Feature**	**Odds Ratio**

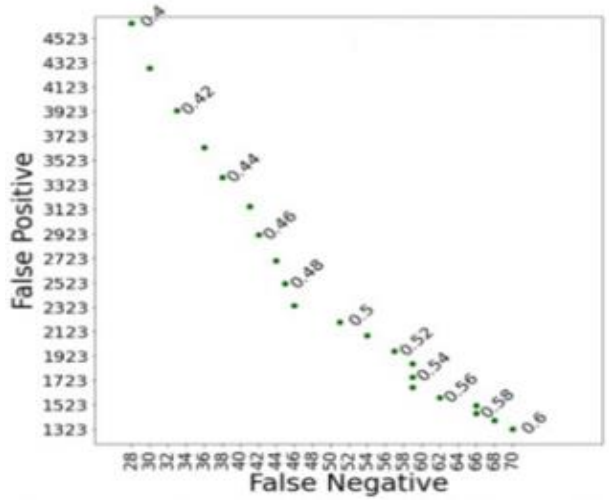
1	Antidiarrheals	2.3529
2	Misc. anti-infectives	1.2558
3	Fluoroquinolones	1.2250
4	Gout agents	1.2108
5	Misc. GI agents	1.1577
6	Penicillins	1.1491
7	Local anesthetics – parenteral	1.0816
8	Unassigned group	1.0391
9	Minerals & electrolytes	1.0265
10	Analgesics – opioids	1.0242
11	Age (per year)	1.0144
12	Antineoplastics	0.9684
13	Anticoagulants	0.9377
14	Diagnostic products	0.9326
15	Anti-rheumatic agents	0.9299
16	Laxatives	0.9268
17	Ophthalmic agents	0.8889
18	Other or mixed races (vs White)	0.8294
19	Tetracyclines	0.7735
20	Toxoids	0.5740

b) Random Forest – Feature Importance

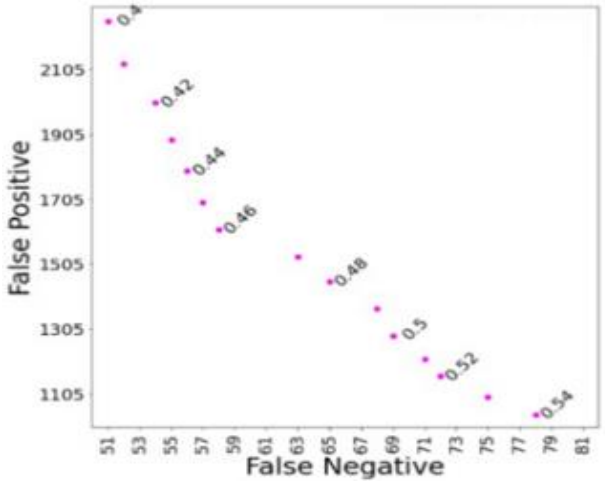
Rank	Feature	Gini Index
1	Minerals & electrolytes	0.1507
2	Misc. anti-infectives	0.1493
3	Unassigned group	0.0686
4	Antiemetics	0.0396
5	Analgesics – opioids	0.0369
6	Diuretics	0.0319
7	Age	0.0295
8	Anticoagulants	0.0261
9	Local anesthetics – parenteral	0.0256
10	Fluoroquinolones	0.0190
11	Assorted classes	0.0190
12	Misc. GI agents	0.0179
13	Antihistamines	0.0172
14	Penicillins	0.0171
15	Corticosteroids	0.0164
16	Ulcer drugs / PPIs	0.0158
17	Hematopoietic agents	0.0149
18	Misc. Hematological agents	0.0147
19	Antineoplastics	0.0144
20	Analgesics – non-opioids	0.0142

Decision boundary analysis

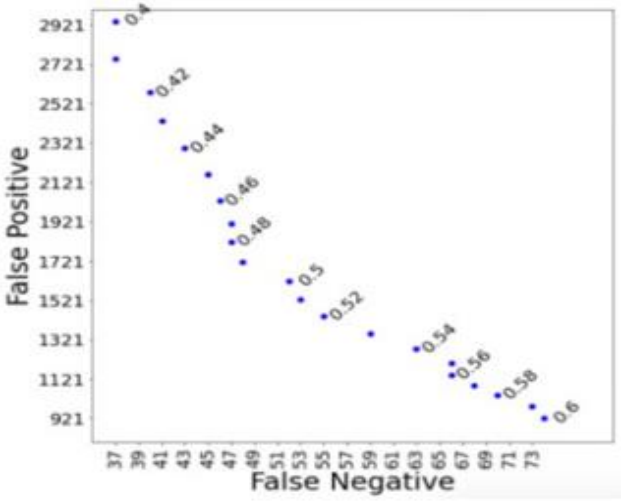
Trade-off curves **(Figure 4)** illustrate how predicted false-negative and false-positive counts shift when the classification threshold is varied from 0.4 to 0.6 in 0.01 increments. Separate curves are shown for LR **(Figure 4a)**, RF **(Figure 4b)**, and the Ensemble model **(Figure 4c)**. Each plotted point represents the pairing of false-negative and false-positive totals on the test set at a specific threshold.



a)



b)



c)

Figure 4. Relationship between false-negative and false-positive predictions as the threshold is adjusted between 0.4 and 0.6 for (a) LR, (b) RF, and (c) Ensemble models.

Findings

All three modeling strategies produced consistently high AUROC values in both cross-validation and test evaluations, indicating that they are suitable tools for estimating the likelihood of a positive CDI finding. The RF model achieved the strongest cross-validated AUROC, whereas the Ensemble approach ranked highest in the final

test stage. Across models, AUROCs ranged from 0.793 to 0.866, and the Ensemble's stability suggests that LR and RF contribute complementary predictive signals. These values exceed those described in earlier research relying on smaller datasets or larger variable sets (0.75–0.82 AUROC in prior studies [18, 20]).

Low p-value predictors in the LR model included several clinically recognizable indicators of CDI risk: patient age [43], prior use of anti-infective medications such as penicillins and fluoroquinolones [44–46], signs of immune suppression due to cancer therapies and associated treatment-related diarrhea prompting more frequent testing [47], and markers of historical gastrointestinal issues (Misc. GI). Similar importance patterns emerged in the RF model through its Gini rankings. This overlap reinforces the clinical relevance of these predictors. In particular, antibiotic exposure again appears to exert substantial influence, underscoring the need for cautious prescribing. Previous research shows that reducing patient susceptibility—which is heightened by antibiotic-driven colonization and progression to CDI [48]—offers greater benefit for prevention than merely limiting transmission [49].

The threshold–performance curves serve as an easily interpretable reference for clinicians, infection-prevention teams, and laboratory staff when shaping testing strategies. By examining these plots, users can pinpoint cutoff values that curb false negatives while also lowering false positives relative to nearby thresholds. Examples include a threshold of 0.55 for the Logistic Regression model, which cuts roughly 100 false-positive predictions, and thresholds of 0.48 and 0.56 for the Ensemble model, each lowering false positives by about 50–100 cases. These visual tools may also support hospital leadership in conducting economic evaluations [50]. Threshold adjustments can be incorporated into cost projections related to diagnostic revisions [51]. Financially, a CDI-positive inpatient with health plan coverage is estimated to incur approximately $21,000 more in medical expenses than a comparable CDI-negative inpatient [52], with the cost rising further when the infection recurs [53]. Meanwhile, the price for a CDI stool assay lies between $15 and $128 (as of 2021) [54]. When combined with clinical familiarity and operational experience, applying threshold-based modeling can provide valuable guidance for balancing overtreatment risk against underdiagnosis. This principle is also applicable to other high-impact pathogens, where avoiding false negatives is critical even if it means tolerating more false positives—COVID-19 being one notable case [55].

From an operational standpoint, these predictions may influence how care facilities manage patient flow, enhance decontamination efforts, or monitor high-risk individuals more closely. In fact, one recent investigation identified a single CT machine as the source of CDI transmission within a major academic medical center [56], underscoring the importance of environmental vigilance even when spread is not outwardly apparent.

Limitations

The study has several constraints:

1. Choice of modeling strategies and calibration. To emphasize interpretability, the study relied on LR, RF, and an Ensemble method. More complex approaches—such as deep learning architectures (RNNs, transformers), or ensemble variations including boosting [57] and stacking [58]—were not explored. Model calibration metrics like the estimated calibration index (ECI) [59], recalibration techniques such as isotonic regression [60, 61], and alternative class-balancing tactics (e.g., upsampling or downsampling [62]) remain untested.
2. Potential dependencies among predictors. Interactions or correlations between pharmacologic classes were not examined, and additional assessment may clarify how these interdependencies influence CDI susceptibility.
3. Clinical deployment. Although real patient data were used during model construction, integrating these models seamlessly into day-to-day clinical operations requires further study. Additional work is necessary to determine whether these predictions are valid for high-risk individuals who might not typically undergo CDI testing.
4. Generalizability and real-world rollout. Model evaluation thus far has been limited to retrospective UCSD data. The models have not yet been implemented at UCSD itself, nor assessed across other healthcare systems or international environments.
5. Distinguishing colonization from active infection. The models identify positive test outcomes without differentiating between true infection and colonization by C. difficile. Although colonization is strongly correlated with infection, it is not a perfect proxy, and deeper analysis of this relationship is still required.

6. Economic impact**.** The financial implications of modifying threshold cutoffs—particularly costs related to monitoring borderline patients—have not been fully assessed and will require a more detailed economic review.

Conclusion

The machine learning models developed in this study, based on extensive longitudinal medication histories and demographic data, demonstrate high predictive accuracy for identifying a patient's first positive CDI test. The models were trained on a substantial dataset of 157493 UCSD Health patients, incorporating a 3-year observational window and 104 covariates. They produced AUROC values ranging from 0.839 to 0.866, and their outputs highlight clinically recognized risk indicators, including advanced age, antibiotic exposure, cancer-related treatments, and gastrointestinal conditions. The threshold-based analyses further offer clinicians flexibility to tailor their preferred balance between sensitivity and specificity. These concepts also generalize to other infectious threats—such as COVID-19—where the consequences of missed diagnoses can be severe. Given that both CDI and COVID-19 disproportionately affect individuals with comorbidities, early detection supported by such predictive tools may reduce transmission, avoid critical complications, and help lower healthcare expenditures while supporting institutional efforts to meet CDC standards.

Acknowledgments: 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.

Conflict of Interest: None

Financial Support: 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.

Ethics Statement: 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).

References

1. Fordtran JS. Colitis due to Clostridium difficile toxins: underdiagnosed, highly virulent, and nosocomial. Proc (Bayl Univ Med Cent). 2006;19(1):3–12. doi:10.1080/08998280.2006.11928114
2. Rupnik M, Wilcox MH, Gerding DN. Clostridium difficile infection: new developments in epidemiology and pathogenesis. Nat Rev Microbiol. 2009;7(7):526–36. doi:10.1038/nrmicro2164
3. Jump RL, Donskey CJ. Clostridium difficile in the long-term care facility: prevention and management. Curr Geriatr Rep. 2015;4(1):60–9. doi:10.1007/s13670-014-0108-3
4. Centers for Disease Control and Prevention. Clostridioides difficile infection [Internet]. Atlanta (GA): CDC; 2020 [cited 2024]. Available from: https://www.cdc.gov/cdiff/index.html
5. Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2019 [Internet]. Atlanta (GA): CDC; 2019. Available from: https://www.cdc.gov/antimicrobial-resistance/data-research/threats/index.html
6. Centers for Disease Control and Prevention. Current HAI progress report [Internet]. Atlanta (GA): CDC; 2019. Available from: https://www.cdc.gov/hai/data/portal/progress-report.html
7. Alrawashdeh M, Rhee C, Hsu HE, Wang R, Horan K, Lee GM. Assessment of federal value-based incentive programs and in-hospital Clostridioides difficile infection rates. JAMA Netw Open. 2021;4(10):e2132114. doi:10.1001/jamanetworkopen.2021.32114
8. UC San Diego Health. Healthcare-associated infections report [Internet]. San Diego (CA): UCSD Health; 2017. Available from: https://health.ucsd.edu

9. Dyer C, Hutt LP, Burky R, Joshi LT. Biocide resistance and transmission of Clostridium difficile spores spiked onto clinical surfaces from an American health care facility. Appl Environ Microbiol. 2019;85(17):e01090-19. doi:10.1128/AEM.01090-19
10. Edwards AN, Karim ST, Pascual RA, Jowhar LM, Anderson SE, McBride SM. Chemical and stress resistances of Clostridium difficile spores and vegetative cells. Front Microbiol. 2016;7:1698. doi:10.3389/fmicb.2016.01698
11. Rineh A, Kelso MJ, Vatansever F, Tegos GP, Hamblin MR. Clostridium difficile infection: molecular pathogenesis and novel therapeutics. Expert Rev Anti Infect Ther. 2014;12(1):131–50. doi:10.1586/14787210.2014.866515
12. Longtin Y, Paquet-Bolduc B, Gilca R, Garenc C, Fortin E, Longtin J, 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. 2016;176(6):796–804. doi:10.1001/jamainternmed.2016.0177
13. Lee HS, Plechot K, Gohil S, Le J. Clostridium difficile: diagnosis and the consequences of overdiagnosis. Infect Dis Ther. 2021;10(1):1–11. doi:10.1007/s40121-020-00370-3
14. Washington State Department of Health. Clostridium difficile testing guidance [Internet]. Olympia (WA): WA DOH; 2019. Available from: https://doh.wa.gov
15. Zacharioudakis IM, Zervou FN, Pliakos EE, Ziakas PD, Mylonakis E. Colonization with toxinogenic C. difficile upon hospital admission, and risk of infection: a systematic review and meta-analysis. Am J Gastroenterol. 2015;110(3):381–90. doi:10.1038/ajg.2015.22
16. Curry SR, Muto CA, Schlackman JL, Pasculle AW, Shutt KA, Marsh JW, 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. 2013;57(8):1094–102. doi:10.1093/cid/cit475
17. Biswas JS, Patel A, Otter JA, van Kleef E, Goldenberg SD. Contamination of the hospital environment from potential Clostridium difficile excretors without active infection. Infect Control Hosp Epidemiol. 2015;36(8):975–7. doi:10.1017/ice.2015.79
18. Wiens J, Horvitz E, Guttag J. Patient risk stratification for hospital-associated C. difficile as a time-series classification task. Adv Neural Inf Process Syst. 2012;25:1–9.
19. Dubberke ER, Yan Y, Reske KA, Butler AM, Doherty J, Pham V, et al. Development and validation of a Clostridium difficile infection risk prediction model. Infect Control Hosp Epidemiol. 2011;32(4):360–6. doi:10.1086/658944
20. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, 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(4):425–33. doi:10.1017/ice.2018.16
21. Katz DA, Lynch ME, Littenberg B. Clinical prediction rules to optimize cytotoxin testing for Clostridium difficile in hospitalized patients with diarrhea. Am J Med. 1996;100(5):487–95. doi:10.1016/s0002-9343(95)99956-3
22. Teeple A, Kane-Gill S, Barsanti F, Tsui FR. Early prediction of positive Clostridioides difficile test results using clinical notes. AMIA Annu Symp Proc. 2021;2021:1088–97.
23. US Food and Drug Administration. FDA pharmacologic class [Internet]. Silver Spring (MD): FDA; n.d. Available from: https://www.fda.gov
24. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(5):e1379. doi:10.1002/widm.1379
25. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12–8. doi:10.11613/BM.2014.003
26. Nembrini S, König IR, Wright MN. The revival of the Gini importance for random forests. Bioinformatics. 2018;34(21):3711–8. doi:10.1093/bioinformatics/bty373
27. Castelvecchi D. Can we open the black box of AI? Nature. 2016;538(7623):20–3. doi:10.1038/538020a
28. Medsker LR, Jain LC, editors. Recurrent neural networks: design and applications. Boca Raton (FL): CRC Press; 2001.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

30. Weerakody PB, Wong KW, Wang G, Ela W. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing. 2021;441:161–78. doi:10.1016/j.neucom.2021.02.046
31. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72. doi:10.1016/j.jbi.2009.08.007
32. Kuo TT, Kim J, Gabriel RA. Privacy-preserving model learning on a blockchain network-of-networks. J Am Med Inform Assoc. 2020;27(3):343–54. doi:10.1093/jamia/ocz214
33. Li MM, Pham A, Kuo TT. Predicting COVID-19 county-level case trends using demographics and social distancing policies with an autoregressive model. JAMIA Open. 2022;5(3):ooac056. doi:10.1093/jamiaopen/ooac056
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30. Available from: https://scikit-learn.org/stable/
35. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5. Available from: https://matplotlib.org/
36. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. doi:10.1038/s41592-019-0686-2. Available from: https://scipy.org/
37. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005;38(5):404–15. doi:10.1016/j.jbi.2005.02.008
38. Klünemann M, Andrejev S, Blasche S, Mateus A, Phapale P, Devendran S, et al. Bioaccumulation of therapeutic drugs by human gut bacteria. Nature. 2021;597(7877):533–8. doi:10.1038/s41586-021-03891-8
39. Imhann F, Vich Vila A, Bonder MJ, Lopez Manosalva AG, Koonen DPY, Fu J, et al. The influence of proton pump inhibitors and other commonly used medication on the gut microbiota. Gut Microbes. 2017;8(4):351–8. doi:10.1080/19490976.2017.1284732
40. Aldape MJ, Packham AE, Nute DW, Bryant AE, Stevens DL. Effects of ciprofloxacin on the expression and production of exotoxins by Clostridium difficile. J Med Microbiol. 2013;62(Pt 5):741–7. doi:10.1099/jmm.0.056218-0
41. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Adv Neural Inf Process Syst. 2013;26:431–9.
42. Baxter SL, Marks C, Kuo TT, Ohno-Machado L, Weinreb RN. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. Am J Ophthalmol. 2019;208:30–40. doi:10.1016/j.ajo.2019.07.005
43. Pépin J, Valiquette L, Cossette B. Mortality attributable to nosocomial Clostridium difficile-associated disease during an epidemic caused by a hypervirulent strain in Quebec. CMAJ. 2005;173(9):1037–42. doi:10.1503/cmaj.050978
44. Deshpande A, Pasupuleti V, Thota P, Pant C, Rolston DDK, Sferra TJ, et al. Community-associated Clostridium difficile infection and antibiotics: a meta-analysis. J Antimicrob Chemother. 2013;68(9):1951–61. doi:10.1093/jac/dkt129
45. Pépin J, Saheb N, Coulombe MA, Alary ME, Corriveau MP, Authier S, 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. 2005;41(9):1254–60. doi:10.1086/496986
46. Vardakas KZ, Trigkidis KK, Boukouvala E, Falagas ME. Clostridium difficile infection following systemic antibiotic administration in randomised controlled trials: a systematic review and meta-analysis. Int J Antimicrob Agents. 2016;48(1):1–10. doi:10.1016/j.ijantimicag.2016.04.008
47. Leffler DA, Lamont JT. Clostridium difficile infection. N Engl J Med. 2015;372(16):1539–48. doi:10.1056/NEJMra1403772
48. Warn P, Thommes P, Sattar A, Vaddi S, Keighley C, Vickers RJ, et al. Disease progression and resolution in rodent models of Clostridium difficile infection and impact of antitoxin antibodies and vancomycin. Antimicrob Agents Chemother. 2016;60(11):6471–82. doi:10.1128/AAC.00974-16
49. Starr JM, Campbell A, Renshaw E, Poxton IR, Gibson GJ. Spatio-temporal stochastic modelling of Clostridium difficile. J Hosp Infect. 2009;71(1):49–56. doi:10.1016/j.jhin.2008.09.013
50. Garrison LP Jr, Babigumira JB, Masaquel A, Wang BC, Lalla D, Brammer M. 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. 2015;18(4):541–6. doi:10.1016/j.jval.2015.02.009

51. Lafata JE, Simpkins J, Lamerato L, Poisson L, Divine G, Johnson CC. The economic impact of false-positive cancer screens. Cancer Epidemiol Biomarkers Prev. 2004;13(12):2126–32.
52. Zhang S, Palazuelos-Muñoz S, Balsells EM, Nair H, Chit A, Kyaw MH. Cost of hospital management of Clostridium difficile infection in United States — a meta-analysis and modelling study. BMC Infect Dis. 2016;16(1):447. doi:10.1186/s12879-016-1786-6
53. Rodrigues R, Barber GE, Ananthakrishnan AN. A comprehensive study of costs associated with recurrent Clostridium difficile infection. Infect Control Hosp Epidemiol. 2017;38(2):196–202. doi:10.1017/ice.2016.246
54. MDsave. Stool C. difficile test cost [Internet]. MDsave; 2021. Available from: https://www.mdsave.com/procedures/stool-c-diff-test/d787ffc4
55. Dai T, Singh S. Overdiagnosis and undertesting: a model of testing with social dynamics. Mark Sci. 2024. doi:10.1287/mksc.2022.0038
56. Murray SG, Yim JWL, Croci R, Bhatt DL, Bafna V, Fowler VG Jr, et al. Using electronic health records to derive control groups for observational studies of vaccine effectiveness: a case study of Clostridium difficile infection. JAMA Intern Med. 2017;177(12):1863–65. doi:10.1001/jamainternmed.2017.5506
57. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms: from machine learning to statistical modelling. Methods Inf Med. 2014;53(6):419–27. doi:10.3414/ME13-01-0122
58. Džeroski S, Ženko B. Is combining classifiers with stacking better than selecting the best one? Mach Learn. 2004;54(3):255–73. doi:10.1023/B:MACH.0000015881.36452.6e
59. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. doi:10.1016/j.jclinepi.2015.12.005
60. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning; 2005; Bonn, Germany. p. 625–32.
61. Edelson M, Kuo TT. Generalizable prediction of COVID-19 mortality on worldwide patient data. JAMIA Open. 2022;5(2):ooac036. doi:10.1093/jamiaopen/ooac036
62. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: a review. Int J Adv Soft Comput Appl. 2013;5(3):176–204.