An Integrated Pipeline for Prediction of Clostridioides difficile Infection

Jiang Li
Durgesh Chaudhary
Vaibhav Sharma
Vishakha Sharma
Venkatesh Avula

See next page for additional authors

Follow this and additional works at: https://digitalcommons.kansascity.edu/studentpub
An integrated pipeline for prediction of Clostridioides difficile infection

Jiang Li1, Durgesh Chaudhary2,7, Vaibhav Sharma3, Vishakha Sharma4, Venkatesh Avula1, Paddy Ssentongo5, Donna M. Wolk6, Ramin Zand2,7 & Vida Abedi1,5*

With the expansion of electronic health records (EHR)-linked genomic data comes the development of machine learning-enable models. There is a pressing need to develop robust pipelines to evaluate the performance of integrated models and minimize systemic bias. We developed a prediction model of symptomatic Clostridioides difficile infection (CDI) by integrating common EHR-based and genetic risk factors (rs2227306/IL8). Our pipeline includes (1) leveraging phenotyping algorithm to minimize temporal bias, (2) performing simulation studies to determine the predictive power in samples without genetic information, (3) propensity score matching to control for the confoundings, (4) selecting machine learning algorithms to capture complex feature interactions, (5) performing oversampling to address data imbalance, and (6) optimizing models and ensuring proper bias-variance trade-off. We evaluate the performance of prediction models of CDI when including common clinical risk factors and the benefit of incorporating genetic feature(s) into the models. We emphasize the importance of building a robust integrated pipeline to avoid systemic bias and thoroughly evaluating genetic features when integrated into the prediction models in the general population and subgroups.

Abbreviations
AUROC  Area under the receiver operating characteristic
Treebag  Bagging for tree
BMI  Body Mass Index
CXCL  Chemokine (C–X–C motif) ligand
C. difficile  Clostridioides difficile
CDI  Clostridioides difficile Infection
CV  Cross-validation
EHR  Electronic health records
eQTL  Expression quantitative trait loci
xgbDART  Extreme gradient boosting with the dropout regularization for regression trees
GWAS  Genome-wide association study
gbm  Gradient boosting machine
IBD  Inflammatory bowel disease
IL-8  Interleukin-8
ICD  International classification of disease
ML  Machine learning
MHC  Major histocompatibility complex
MAF  Minor allele frequency
MCC  Matthews correlation coefficient
nnet  Neural network
NPV  Negative predictive value
Phenome-wide association study
PPV  Positive predictive value

1Department of Molecular and Functional Genomics, Geisinger Health System, Danville, PA, USA. 2Neuroscience Institute, Geisinger Health System, Danville, PA, USA. 3Geisinger Commonwealth School of Medicine, Danville, PA, USA. 4College of Osteopathic Medicine, Kansas City University, Kansas City, MO, USA. 5Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA, USA. 6Molecular and Microbial Diagnostics and Development, Geisinger Medical Center, Danville, PA, USA. 7Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA. *email: vidaabedi@gmail.com
With biobank and genetic data integrated with electronic health records (EHR) comes the development of predictive models designed for healthcare applications. There is an urgent need to develop robust modeling pipelines using machine learning (ML) to determine whether EHR-derived common clinical risk factors can predict the phenotype of interest and whether adding genetic factors can improve model performance. Several factors have to be considered during the model development, including (1) selection bias for the biobank data; (2) case–control imbalance; (3) temporal bias in feature acquisition; (4) impact of confounding factors; (5) optimal model selection to capture multi-way interactions; and (6) predictive power and generalizability of the final models.

Here we choose *Clostridioides difficile* infection (CDI) as a proof-of-concept given its complexity and clinical importance. The purpose of this study is to develop an integrated pipeline for predicting symptomatic CDI using common EHR-derived clinical and genetic risk factors. Focusing on symptomatic CDI is driven by the fact that testing and treatment for CDI are not recommended in asymptomatic individuals.

CDI is considered the most common cause of healthcare-associated diarrhea and is listed as one of the top five urgent antimicrobial resistance threats by the Centers for Disease Control and Prevention (https://www.cdc.gov/drugresistance/biggest-threats.html). Existing literature reporting CDI prediction focuses on three outcomes—symptomatic infection, severity of the infection, and recurrence. The prediction models developed in this field vary by setting, patient recruitment, data source, study design, feature selection, and algorithms. EHR-based studies have become popular due to their improved predictability, specificity, and generalizability. Host genetic susceptibility to CDI and epidemiology of *C. difficile* strains have been topics of investigations. Intestinal inflammatory cytokines correlate more closely to disease severity than pathogen burden. The same mechanism (inflammatory cytokines) applies to the inflammatory cytokine signature (plasma level of IL-6, IL-8, and TNF-α) for the prediction of COVID-19 severity and survival. A previous candidate gene approach revealed that genetic polymorphisms, rs4076(–251T>A) or rs2227306(+ 781T/C) from a pro-inflammatory cytokine, are associated with IL-8 production and predisposition to CDI with functional impact (eFigure 1). Genetic markers are not yet included in any established disease scoring system or clinical decision tool for risk stratification of CDI due to unclear causality, small effect size, complex gene by environment (GxE) interaction, and data availability/accessibility.

**Materials and methods**

The study was conducted and reported according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guideline. The Geisinger Institutional Review Board approved this study to meet “Non-human subject research” using de-identified information. All research was performed in accordance with relevant guidelines/regulations. Geisinger built and performs regular updates to the de-identified structured EHR database for research linked to the MyCode Community Health Initiative biorepository. The study was conducted and reported according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guideline.

**Robust phenotyping algorithm**

The phenotype algorithm used to identify CDI cases and controls from EHR data was developed by eMERGE entitled “Phenotype Algorithm Pseudo Code (August 16, 2012)” and collected at PhenoKB (https://phekb.org/phenotype/clostridium-difficile-colitis/). This algorithm adopted the golden standard, which used laboratory test data to identify CDI cases. Based on clinical symptoms, we first identified adults (age ≥ 18) with CDI from the Geisinger EHR. Three or more consecutive liquid stools within a day may be tested for *C. difficile* based on recommended guidelines. Patients tested for *C. difficile* but with negative results or exposed to antibiotics with high or moderate risk (Appendix-1) served as controls. Nongenetic risk factors collected in this study included age index, people in healthcare settings, antibiotic treatment, underlying comorbidities such as inflammatory bowel disease (IBD), type 2 diabetic mellitus (T2DM), HIV, cancer, medications such as chemo, transplant, corticosteroid, anti-TNF, proton pump inhibitors (PPI). The observation window for each risk factor was empirically defined to avoid an uneven sampling of the disease trajectory, which might lead to temporal bias in feature selection. Clinical risk factors and demographic information were extracted from the structured EHR based on the International Classification of Disease (ICD)-related codes and medication codes (Appendix 1).

Using data from January 1, 2009, through December 31, 2017, we identified 6,035 cases and 72,241 controls. Of these, 5911 cases and 69,086 controls had self-reported European ancestry. Overall, 22.4% (1156/14,148 for case/control) of European (EUR) patients enrolled in the MyCode project with genetic data available.

**Data pre-processing**

The entire cohort of participants with EUR ancestry (n = 5911/69,086) was first split based on the availability of genetic data. No missing data was observed in any of the included clinical variables. The MyCode samples were genotyped as previously reported. Both SNPs (rs2227306 and rs4076) from IL-8 passed the quality control.
without missingness. The genetic variable is routinely treated in a dosage manner (0, 1, 2). Data extraction and pre-processing details (z-scored index age, binary codes for other variables) have previously been described.

Either the $\chi^2$ test, substituted by Fisher exact test for frequency per group ≤ 5% or the ANOVA test was performed to screen for the bivariate association. A heatmap was created to show the significance of this bivariate association based on the log-transformed p-value from each bivariate $\chi^2$ test (Fig. 2).

Simulations to determine predictive power
A genotype simulation study was conducted to determine different modeling algorithms’ predictive power and generalizability in the nonMyCode cohort. The genotypes of rs2227306 (IL-8) in this cohort of 4755 cases and 54,938 controls were created by a simulation strategy based on an assumption of a binomial probability distribution of each allele equaling to the prior parameter, MAF, estimated from the corresponding MyCode subgroups stratified by CDI, age, and sex (Table 2).

$$\text{AlleleDosage}_{i \in R} = \text{rbinom}(N_{i \in R}, 1, \text{MAF}_{i \in R}) + \text{rbinom}(N_{i \in R}, 1, \text{MAF}_{i \in R}).$$

where $R$ is a vector or a matrix of the summary statistics derived from the MyCode cohort, $N$ and $\text{MAF}$ represent the number of subjects and the corresponding MAF for each subcategory. This simulation strategy considers confounding factors such as age and sex, which may impact the association between the genetic variant and the outcome variable.

Individual SNP association testing
Genotype and phenotype association was conducted using Logistic Regression after controlling covariates such as age or sex in subgroups stratified by sex or age (binary).

Controlling confounding
Age and sex were identified as confounding factors for the association between genetic/nongenetic risk factors and CDI (see “Result”). They were selected as covariates in a Logistic Regression model to create propensity scores (R MatchIt package). We chose “nearest neighbor matching without replacement” to create a more balanced case:control ratio at 1:5 or 1:10, as shown in eFigure 2.
Figure 2. Association among features and performance of models in prediction of CDI in MyCode and nonMyCode samples with (simulated) genotypes included. (a, b) Heatmaps to show the significant association between variables employed in the prediction model using the training dataset. Data extraction and preprocessing details (z-scored index age, binary codes for other variables) have previously been described. Association among variables (index age further dummy coded) was assessed using a bivariate $\chi^2$ test. (c, d) To examine the discrimination power of each modeling algorithm in the testing dataset, we estimated the AUROC using common clinical risk factors for CDI with or without rs2227306 as predictors. Here the genotypes of rs2227306 were simulated in the nonMyCode samples. (e) The summary of AUROCs of the optimal modeling algorithms (gbm and xgbDART) versus glm using simulated rs2227306 genotype. P values represent the result of the DeLong test to compare AUROC between models with or without (simulated) genetic data included, with or without PSM for index age and sex.
Selecting machine learning (ML) algorithms
Model development and optimization were based on selecting the optimal sampling approach followed by a comparison of the eight classification algorithms, including Logistic Regression (glm), Gradient Boosted Classification (gbm), Extreme Gradient Boosting with the dropout regularization for regression trees (xgbDART), Bagging for tree (treebag), Neural Network [nnet, using 1-layer fully connected neural networks (shallow)], C5.0 (c5), LogitBoost (lb), and Support Vector Machine (SVM). We selected these models based on model/algorithmic diversity, performance on similar tasks, model interpretability, and ease of implementation. Specifically, these set of algorithms are well-established with good performance for tabular data in a wide range of classification tasks. Furthermore, Ensemble methods (C5.0, xgbDART, Bagging, LogitBoost) that combine multiple weak learners to create a stronger model, can lead to improved predictive performance and better generalizability. Finally, by creating decision boundaries that are complex and not linear, Neural Network and SVM are capable of capturing non-linear relationships between the variables.

Evaluation metrics
We assessed the performance of the multivariable model in the prediction of CDI mainly by sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), precision, recall, F1 score, Matthews Correlation Coefficient (MCC), accuracy, and Area Under the curve (AUROC) for a classification problem.

\[
F1 \text{ Score} = 2 \ast (\text{Recall} \ast \text{Precision}) / (\text{Recall} + \text{Precision})
\]

\[
\text{MCC} = (\text{TP} \ast \text{TN} - \text{FP} \ast \text{FN})/\sqrt{\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}
\]

where TP = True positives; TN = True negatives; FP = False positives; FN = False negatives. Recall = TP/(TP + FN); Precision = TP/(TP + FP). Both F1 Score and MCC are calculated based on the confusion matrix and are reliable metrics for evaluating binary classification models, particularly in imbalanced datasets. F1 score combines precision and recall to provide a balanced evaluation of a model's accuracy. MCC considers the overall balance between the four elements of the confusion matrix. The value of MCC ranges from -1 to 1, where 1 or -1 indicates the complete agreement or disagreement between predicted classes and actual classes, respectively, and 0 indicates completely random guessing. The result of these metrics for the training and testing datasets across eight ML algorithms was summarized in eTable3. AUROC is a graphical representation of the model's TP rate (sensitivity/recall) versus the FP rate (1-specificity) at various probability thresholds. AUROC is insensitive to the class distribution and is often preferred over accuracy for imbalanced datasets.

Addressing data imbalance
The imbalance of a case:control dataset can lead to biased model training, where the ML models tend to favor the majority class and underestimate the minority class, resulting in high accuracy by always predicting the majority class. The oversampled data improve ML models for the prediction of minority class when there is a significant unbalanced case:control cohort, as shown in this study (\~{} 1:12). We performed oversampling (using Synthetic Minority Oversampling Technique, SMOTE, and random oversampling, ROSE) of the minority class during the model training to address the class imbalance and used F1 score and MCC to assess the model performance.

The SMOTE function oversampled the minority class (rare events) using bootstrapping (perc.over = 100) and k-nearest neighbor (k = 5) to synthetically create additional observations of that event and undersampling the majority class (perc.under = 200). For each case in the original dataset belonging to the minority class, perc. over/100 new examples of that class were created. ROSE applied smoothed bootstrapping to draw artificial samples from the feature space around the minority class without undersampling the majority class. We oversampled the minority class to reach a case:control ratio of 1:1 in a training dataset.

Model optimization
Both MyCode and nonMyCode samples were split into training and holdout data(testing) with 7:3 ratio at the beginning of this study. All the models were tuned in training data and tested in holdout data(testing data), as described in Fig. 1. The fivefold repeated CV has been applied to the corresponding training data from either MyCode or nonMyCode cohorts. The nonMyCode samples were considered as an additional dataset to determine the overfitting. However, the genotypes for nonMyCode samples were not available, which is very common for the non-biobank population. That is why we impute the genotype data for nonMyCode samples by simulation to determine the generalizability of the optimal models. A hyperparameter tuning grid was used to train the model with five-fold repeated cross-validation (CV) and ten repeats (R caret package). Model tuning was performed by an automatic grid search for each algorithm parameter randomly (eTable1 for the tuning parameters used in each final model). Finally, the testing set was used to calculate the model AUROC.

We also compared optimal models with the benchmark algorithm glm, and extracted the feature importance of the 12 included variables. The ranks of each variable in feature importance, particularly the genetic variable, were compared across the algorithms. The DeLong test was used to compare AUROCs from two modeling algorithms and compare AUROCs of the best model with or without the genetic feature included. For each model, the AUROC was calculated first, and the 95% confidence interval (95% CI) of AUROC was also computed. The covariance matrix \( \text{Cov}(\hat{A}, \hat{B}) \), where \( A \) and \( B \) were the estimated AUROC for models A and B, respectively, of the AUROCs between the two models was calculated. The formula for the covariance between two sample
proportions, $\hat{A}$ and $\hat{B}$, was given by: $\text{Cov}(\hat{A}, \hat{B}) = \left(1 - \frac{\hat{A}}{n_A}\right) + \left(1 - \frac{\hat{B}}{n_B}\right) + \frac{\hat{A}}{n_A} + \frac{\hat{B}}{n_B} - 1$, where $n_A$ and $n_B$ were the sample sizes for models A and B, respectively. The Z statistic $\frac{\hat{A} - \hat{B}}{\text{SE}(\hat{A} - \hat{B})}$, where $\text{SE}(\hat{A} - \hat{B}) = \sqrt{\text{Cov}(\hat{A}, \hat{B})}$, and p-value were computed under an assumption of normal distribution. This p-value represented the probability of observing a difference in AUROC as extreme as the one observed in the data, assuming the null hypothesis where there was no difference between the two models.

**Results**

**Patients demographics**

The entire dataset was split based on the availability of the genetic data. Demographic and clinical information for MyCode and nonMyCode cohorts are listed in Table 1. The case:control ratio in MyCode (n = 1156/14,148, 1:11.55) was comparable to that in the nonMyCode cohort (n = 4755/54,938, 1:12.24). This ratio (~1:12) indicated approximately a ten-fold enrichment for controls as shown in the previously reported EHR-based large populational studies (more than 1:100). Several demographic features (e.g., sex, age) and known clinical risk factors showed significant differences between case and control groups in MyCode and nonMyCode cohort. Their bi-variate association among all predictive variables is illustrated in Fig. 2a and b. Antibiotics were the most significant risk factor for CDI.

**Genotype–phenotype association**

The Logistic Regression analysis showed a significant association between rs2227306 genotype and CDI only in young MyCode patients ($\beta = 0.138$, p = 0.048 vs. $\hat{\beta} = 0.062$, p = 0.263) after controlling for sex. After controlling for age, this nominal association was only observed in females ($\beta = 0.119$, p = 0.034 vs $\hat{\beta} = 0.053$, p = 0.427). The minor alleles from both SNPs with the higher expression level of CXCL8 and CXCL6 were associated with an increased risk for CDI.

**Comparing oversampling methods to manage the case–control imbalance**

SMOTE outperformed ROSE in seven out of eight examined algorithms (Fig. 3). When using xgbDART and gbm models, SMOTE in the training dataset led to better F1 (0.264 and 0.272, respectively) than the ROSE (0.253 and 0.261, respectively) in the testing dataset. The process without resampling provided the worst F1 (0.037 and 0.056, respectively). SMOTE was chosen for the following analyses. Both MCC and F1 scores showed that SMOTE yielded a slightly higher value than ROSE with or without the genetic feature included according to xgbDART and gbm models in the testing dataset (eTable 2).

**Predicting CDI in MyCode patients with or without propensity score matching (PSM)**

The best tuning parameters selected for each model based on the largest ROC value from the training dataset were summarized in eTable 1. Among eight algorithms examined, we found that using the 11 clinical risk factors with rs2227306, in conjunction with gbm and xgbDART led to superior results (with the genetic feature, AUROC$_{\text{gbm}} = 0.72$[0.694–0.746] versus AUROC$_{\text{glm}} = 0.684[0.655–0.713]$, p = 1.92e−07) (Fig. 2c). There was no significant difference between gbm and xgbDART (AUROC$_{\text{xgbDART}} = 0.715[0.689–0.742]$, p = 0.247). The genetic feature was always ranked higher in gbm and xgbDART than glm (Fig. 4 the top two rows). The result of performance comparison with rs2227306, in conjunction with gbm and xgbDART led to superior results (with the genetic feature, AUROC$_{\text{gbm}} = 0.820[0.809–0.831]$ and AUROC$_{\text{xgbDART}} = 0.819[0.808–0.831]$) outperformed other modeling algorithms (e.g., AUROC$_{\text{glm}} = 0.751[0.737–0.765]$, p = 1.68e−57) (Fig. 2d). Again, rs2227306 was ranked higher in gbm and xgbDART than glm (Fig. 4 the bottom two rows). The result of performance metrics (sensitivity, specificity, PPV, NPV, precision, recall, F1 score, and accuracy) for the prediction of CDI in the training and testing datasets across eight ML algorithms using SMOTE as the oversampling method was summarized in eTable 3.

**Predictive power in nonMyCode patients**

The simulated genotype in the nonMyCode cohort can recapitulate the nominal association between rs2227306 and CDI (Table 2). Consistent with the result from MyCode patients, with the simulated genetic feature included, gbm (AUROC$_{\text{glm}} = 0.820[0.809–0.831]$) and xgbDART (AUROC$_{\text{xgbDART}} = 0.819[0.808–0.831]$) outperformed other modeling algorithms (e.g., AUROC$_{\text{glm}} = 0.751[0.737–0.765]$, p = 1.68e−57) (Fig. 2d). Again, rs2227306 was ranked higher in gbm and xgbDART than glm (Fig. 4 the bottom two rows).
| Data source | Non-MyCode sample (simulated genotype) | MyCode sample (original) |
|-------------|---------------------------------------|--------------------------|
| **Input features** | w/o CDIFF | P_value* | w/o CDIFF | P_value* | w/o CDIFF | P_value* | w/o CDIFF | P_value* | w/o CDIFF | P_value* |
| Sample size | 47,550 | 13,165 | 23,775 | 47,550 | 1156 | 14,148 | 5780 | 11,560 | 229 | 408 |
| **Age at index** | 57.9 (23.9) | 9.31E−28 | 45.5 (26.9) | 1.0 | 2.05E−16 | 0.294 | 2.05E−16 | 0.123 | 2.05E−16 | 0.454 |
| **Sex** | Female | 0.005 | 0.002 | 0.027 | 0.002 | 0.003 | 0.001 | 0.006 | 0.007 | 0.008 |
| **rs2227306** | CC | 0.001 | 0.002 | 0.003 | 0.001 | 0.002 | 0.001 | 0.002 | 0.003 | 0.001 |
| | TT | 0.001 | 0.002 | 0.003 | 0.001 | 0.002 | 0.001 | 0.002 | 0.003 | 0.001 |
| **Antibiotics** | 2007 (0.422) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Inflammatory bowel dis.** | 334 (0.070) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Proton pump inhibitor** | 1365 (0.287) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Corticosteroid** | 1075 (0.226) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Chemotherapy** | 298 (0.063) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Type 2 diabetic melittus** | 1222 (0.257) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **anti-TNF medication** | 19 (0.004) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Transplant medication** | 126 (0.027) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **HIV** | 4 (0.001) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

**Table 1.** Demographic, potential risk factors, and genotype data for MyCode and non-MyCode participants with or without propensity score matching (PSM). Data were presented as the number of subjects with frequency in parentheses or Mean ± SD. All patients were subjected to a molecular test for the detection of *C. difficile*. A *"* represented statistics from the Chi-square test to determine whether there is a significant difference between the expected frequency and observed frequency in one or more categories or statistics from the ANOVA test to determine whether there is a significant difference between group means.

**MHC** major histocompatibility complex, w/ with, w/o without.
Discussion
We developed a prediction model of symptomatic CDI by integrating common risk factors extracted from electronic health records and genetic risk factors (rs2227306/IL8). Our modeling pipeline included steps to minimize systemic bias in the final models while adhering to best practices to improve model transparency. These steps included (1) applying robust and validated phenotyping, (2) selecting and optimizing a range of ML algorithms with a focus on attributes such as generalizability, interpretability, potential interactions, and bias-variance trade-off, (3) addressing data imbalance and performing extensive simulation studies to determine the predictive power in samples without genetic information, and finally using PSM to control for confounding factors.

Overall, our results supported that decision tree-based Ensemble methods such as gbm and xgbDART demonstrated superior discriminative power than glm(logistic regression)\textsuperscript{16}. Both gbm and xgbDART are based on the boosting algorithm\textsuperscript{17,18} which is an ensemble technique that sequentially builds multiple weak learners, each focusing on the mistakes of its predecessor and assigning higher weights to misclassified instances to correct them in subsequent iterations. Ensemble methods often work better than individual ML algorithms due to the following reasons\textsuperscript{17,18}: (1) Handing imbalanced data by giving more weight to the minority class and balancing the predictions as shown in this study; (2) Reduced Bias and Variance by combing multiple individual models(weak learners) to create a single and more robust model(strong learners); (3) Complementary learning by leveraging the complementary strengths of individual algorithms with less overfitting and improved generalization; (4) Tackling complex relationships which cannot be effectively captured by a single model; (5) Robustness to noisy data because of the averaging or voting mechanisms which often are less affected by noisy data points and outliers.

The improvements gained by including common genetic variants in the optimal models were limited and age- or sex-dependent after PSM. This finding was consistent with the decreased genetic heritability observed.
Figure 4. Feature importance for the cohort with or without (simulated) genetic data was plotted for two selected models (gbm and xgbDART), which outperformed other models (glm and nnet). This study was based on 12 features, including one genetic risk factor, rs2227036, from IL8. Feature importance from glm and nnet was always plotted as a control to compare the rank of the features weighted by optimal modeling algorithms (gbm and xgbDART) in MyCode (top two rows) and nonMyCode samples (bottom two rows). The genetic feature was weighted the top tier in gbm and xgbDART but not in glm and nnet irrespective of PSM in the MyCode cohort.
Table 2. Two simulation strategies to create the genetic data of rs2227306(IL8) in the nonMyCode cohort. Significant values are in italics. This table summarizes the sample size and minor allele frequency (MAF) of each subgroup stratified by the outcome variable (CDI), age, and sex in both MyCode and nonMyCode cohorts. The genotype of rs2227306 was simulated based on an assumption of binomial distribution of each allele with frequency equaling the prior parameter (MAF) estimated from the corresponding MyCode samples. The genotype for each subject would be a combination of sampling from two binomial distributions with the same MAF.

| CDI (outcome) | AGE [escore ≥ 0] | SEX (female) | Number of subjects | Minor allele frequency | χ²/p value |
|----------------|------------------|--------------|---------------------|------------------------|------------|
|                | MyCode | NonMyCode | MyCode | NonMyCode | MyCode | NonMyCode | MyCode | NonMyCode |
| 0 | 0 | 0 | 1646 | 12,757 | 0.4067 | 0.4111 | 7.443/0.024 | 4.605/0.1 |
| 0 | 0 | 1 | 4928 | 16,072 | 0.4074 | 0.4098 | | |
| 0 | 1 | 0 | 3553 | 12,066 | 0.4191 | 0.4174 | | |
| 0 | 1 | 1 | 4021 | 14,043 | 0.4193 | 0.4233 | | |
| 1 | 0 | 0 | 157 | 527 | 0.4522 | 0.4639 | 3.515/0.173 | 11.88/0.003 |
| 1 | 0 | 1 | 286 | 688 | 0.4353 | 0.4317 | 4.701/0.095 | 3.009/0.222 |
| 1 | 1 | 0 | 322 | 1464 | 0.4161 | 0.3924 | 0.026/0.987 | 7.23/0.027 |
| 1 | 1 | 1 | 391 | 2076 | 0.4501 | 0.4381 | 3.154/0.207 | 5.37/0.068 |

in late-onset compared to early-onset in multiple complex disease traits, and rs2227306 (IL-8) was related to early-onset of disease.

Importance of genetic risk factor across modeling algorithms
Feature importance was considered a measure of the individual contribution of the feature for a particular classifier, regardless of the shape or direction of the feature effect. The genetic feature was consistently ranked high in the optimal models, even higher in PSM subgroups. This finding corroborates the clinical value of genetic information in prediction models. The conclusion made from the MyCode sample could partially be extended to the larger nonMyCode sample. The potential interaction between the genetic feature and clinical risk factors beyond age and sex was not recognized and implemented in the genotype simulation, which may prevent the conclusion made from the MyCode sample could fully be extended to the larger nonMyCode sample, leading to more uncertainty of the discriminative power in the model with the simulated genetic feature included. The importance of the genetic feature was ranked lowest in the Logistic Regression-based model, suggesting Logistic Regression may underestimate the contribution of the genetic variant for the prediction of CDI, highlighting the importance of capturing multi-way interactions when assessing the value of common genetic variants with a small effect size in prediction models.

Clinical relevance of the selected genetic variant from IL-8
Previous candidate gene approach of genetic predisposition of CDI has revealed a promoter polymorphism rs4073 (−251T>A) and its linkage disequilibrium SNP, rs2227306 (+781T>C), from a pro-inflammatory cytokine, IL-8—can result in increased IL-8 production and predisposes subjects to CDI, recurrent CDI, or severity of CDI. They are both eQTL for a gene cluster located at 4q13.3 (eFigure 1), encoding several members of the CXCL chemokine family such as CXCL8 (aka IL-8), CXCL6, CXCL5, CXCL1, and more, which promote the recruitment of neutrophils to the site of infection.

Minor alleles from both SNPs are associated with an increased expression level of CXCL8 (GTEx). Increased IL-8 protein levels and CXCL5 and IL8 message levels have been associated with prolonged disease. Intracolonic administration of TcdA/TcdB in a mouse model increases the expression of inflammatory mediators such as CXCL1, the murine ortholog of human IL-8, PheWAS analyses from UKBIOBANK (eFigure 1) confirmed that the top phenotypes associated with rs2227306 included some laboratory variables such as “Neutrophil count”, “Neutrophil percentage”, “White blood cell count”, the latter of which contributed to a composite risk score developed in earlier studies in the prediction of CDI severity. Our study is the first to evaluate the integrated model by including the genetic and common clinical risk factors using various optimized ML algorithms to predict CDI.

In the context of other similar studies
Since the testing and treatment are not recommended in asymptomatic carriers of C. difficile, we direct our focus on symptomatic CDI. The results from this study may eventually facilitate at-risk patient stratification for targeted treatment in patients more likely to benefit from emerging prevention or treatment options such as a vaccine, fidaxomicin, monoclonal antibodies, and fecal microbiota transplantation. Results may also support more granular stratification for therapeutic trials.

Controls were defined as patients without CDI, based on negative molecular laboratory results or exposed to similar risk factors, such as antibiotic use. Because of using the eMERGE phenotype algorithm with the inclusion/exclusion criteria to define controls, our case/control ratio was approximately 1:10, which was a tenfold enrichment of controls with increased risk for CDI, compared to more than 1:100 ratio summarized from other retrospective cohorts studies. This enrichment would make the prediction more challenging.
| Study design | Study characteristics (timeline, sample size, and features) | Outcome measures | Handling missing data | Algorithms | Performance | Summary |
|--------------|-------------------------------------------------|-----------------|---------------------|------------|-------------|---------|
| Retrospective cohort\(3^{rd}\) (2021) | EHR (May 2010 to July 2014) from single healthcare; 9986 CDI cases and 2,230,354 members without CDI; 104,518 hospital discharges for validation; case:control = 1:23 = 20 Risk factors | CDI | Not addressed | Univariate logistic regression | Model 1: using hospital discharge as the IDRSA, C-statistic of 0.848 in subsequent 31–365 days; Model 2: using a random date as the IDRSA, C-statistic 0.722 | Identification of high-risk populations for CDI vaccine trials to determine the study feasibility (sample size and time to completion) |
| Case–control study\(3^{rd}\) (2019) | Adult patients admitted to a multicenter study from July 1, 2015, to July 1, 2017, who received systemic antibiotics 200 subjects (100 cases and 100 controls) Reported 2 features | CDI (hospital-associated) | Not addressed | Univariate logistic regression Multivariable logistic regression to formulate a point-based risk prediction model | Sensitivity and specificity were 76% and 49%; Highest accuracy (63%); AUROC = 0.7 | A simple-to-implement hospital-onset CDI risk model; including only independent risks that can be obtained immediately on presentation to the healthcare facility |
| Retrospective Cohort\(4^{th}\) (2018) | EHR-based adult inpatients admitted to two healthcare systems one (January 1, 2010, and January 1, 2014) and the other (June 1, 2012, and June 1, 2014) 191,014 (155,009/36,005 for training/testing) and 65,718 (33,477/32,241 for training/testing) for two healthcare systems respectively Case:control = 1:100 4836 and 1837 features from the two healthcare systems respectively | CDI | Not addressed | 1.2 regularized logistic regression Logistic regression to create a daily risk score for risk stratification | AUROC = 0.82 [0.80–0.84] and 0.75 [0.73–0.78] for two cohorts respectively | Many of the top predictive factors differed between the cohorts from two healthcare systems Institution-specific models instead of “one-size-fits-all” models |
| Retrospective Cohort\(5^{th}\) (2016) | Population-based sample Medicare beneficiaries aged 65 and older on January 1, 2008, with continuous Medicare coverage from January 1, 2008, through December 31, 2009. Inpatient setting (58.5%) Of 1,165,165 Medicare beneficiaries meeting the enrollment criteria, 6,838 had an incident CDI episode; case:control = 1:170 22 features | CDI | Not addressed | Logistic regression model for feature selection sequentially removes features with < 0.8 change in C-statistic; A weighted score was developed for each of the risk factors based on its odds ratio, with the sum of all of the risk values representing a participant’s risk score | C-statistic = 0.858 NPV = 98.7% | Developed a risk stratification scoring system Emphasized the age-dependent CDI |
| Retrospective cohort\(6^{th}\) (2016) | Admitted over a 1-year period (2013) Total of 61,482 subjects, Discovery dataset (40,960) and validation dataset (20,492) case:control = 1:200 ~ 25 features | CDI (hospital-associated) | Not addressed | Multivariable analysis to identify risk factors individually Multivariable model based on six risk factors to develop a risk score | Sensitivity = 82.0%; Specificity = 75.7%; AUROC = 0.85 | Developed a clinical prediction rule to identify patients at high risk for primary CDI |
| Retrospective cohort (longitudinal)\(7^{th}\) (2015) | Hospital discharge data and pharmacy data from two large academic centers linked to active population-based CDI surveillance data from the Emerging Infections Program (EIP) Of the 35,186 index hospitalizations, 288 (0.82%) had CDI ≥ 28 days post discharge 39 features to begin with, 4 features left | CDI (Having CDI ≥ 28 days post discharge) | Not addressed | Cox proportional hazards model (stepwise backward selection) for low and high risk groups | C-statistic = 0.75 | Develop a risk score applied at discharge to identify a risk of CDI ≥ 28 days post discharge |

Continued
| Study design                  | Study characteristics (timeline, sample size, and features)                                                                                                                                                                                                 | Outcome measures                | Handling missing data | Algorithms                           | Performance                                      | Summary                                                                                       |
|------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|-----------------------|--------------------------------------|------------------------------------------------|------------------------------------------------------------------------------------------------|
| Case–control study³⁵         | Patients admitted between January 2005 and December 2011 from a single healthcare system Discovery: 180 cases and 330 controls; Validation: 97 cases and 417 controls; case:control = 1:120 12 features                    | CDI (hospital-associated)        | Not addressed          | Stepwise backward elimination to determine the best fit model Logistic regression to develop a simplified risk score | Corrected AUROC = 0.81 [0.77–0.85]; Calibration: Brier score = 0.004 | Developed and validated a model to predict the incident CDI in hospitalized patients who receive systemic antibiotic treatment |
| Retrospective cohort³⁴       | All patients admitted on or after April 12, 2011 and discharged on or before April 12 Training: 34,846 admissions (372 cases of CDI); Validation: 34,722 admissions (355 cases of CDI); Case:Control = 1:100 14 features (EHR Model) 1017 features (Curated Model); 10,859 features (EHR ALL) | CDI (hours from the time of admission) | Not addressed          | L2-regularized Logistic Regression 3 Models Compared based on different number of features included in the final models to discriminate low-risk from high-risk patients | AUROC Risk Period > 24 h EHR = 0.81 [0.79–0.83] Curated = 0.72 (0.69–0.75) EHR ALL = 0.84 [0.80–0.88] Risk Period > 48 h EHR = 0.7886 (0.76–0.82) Curated = 0.69 (0.66–0.72) EHR ALL = 0.79 (0.76–0.81) | Additional features from EHR data improved prediction and outperformed the model only considering a small set of known clinical risk factors |
| Retrospective cohort³⁴       | All inpatient visits for the 2 years between April 2011 and April 2013 1348 test positive cases of C difficile out of 132,853 admissions from three hospitals, varying in size and location; case:control = 1:100 578 binary features; Different feature spaces including common (256) and specific features | CDI (hospital-associated)        | Missingness has been discussed; source feature space and target feature space | L2-regularized logistic regression Multivariate Logistic regression | AUROC = 0.80 varied by the approach and target task | The external data from other hospitals can be successfully and efficiently incorporated into hospital-specific models |
| Case–control study³⁴         | Not available (abstract only) 8 Known risk factors                                                                                                                                                                                                                             | CDI                              | Not addressed          | All feature included Multivariate regression model to create a weighted score tool | Sensitivity = 92%; Specificity = 59% | Developed a weighted scoring tool to predict incident CDI |
| Retrospective cohort³⁴       | A consecutive cohort of patients admitted to the adult medical service over a period of 17 months (June 2011 to October 2012) 62 out of 7026 patients with over 48 h hospital stay having hospital-onset CDI cases; case:control = 1:100 Reported 6 features | CDI (hospital-onset)             | Addressed for missingness in serum albumin level | Univariate analysis to determine the potential risk factors included in the model Multivariable logistic regression model using a forward stepwise selection for features | AUROC = 0.94 [0.92–0.95]; Sensitivity = 98.3% [90.2–99.9]; Specificity = 85.2% [84.3–86.0] | Developed a predictive scale for hospital-onset CDI which can be used for risk stratification |
| Retrospective Cohort³²       | Patients admitted for ≥ 48 h during the calendar year 2003 from a single healthcare system 35,350 total admissions & 329 CDI cases. Case:control = 1:100 11 features                                                                                     | CDI                              | Not addressed          | Feature selection based on high dimensional data reduction techniques such as PCA, cluster analyses Logistic stepwise regression to determine the best fit model Logistic regression also tests for some feature interactions | C index = 0.88; Brier score 0.009) | Developed and validated a CDI risk prediction model using EHR with strong discriminative capacity |
| Retrospective cohort³¹       | Three phases design: discovery dataset (NA), testing dataset (n = 1,468), and external validation (n = 29,425)                                                                                                                                             | CDI (Clostridium difficile-associated disease (CDAD)) | Not addressed          | Logistic regression model | AUROC = 0.827; Sensitivity = 70% and specificity = 95% | Developed a predictive score to predict patients’ risk of developing CDAD |

Continued
| Study design | Study characteristics (timeline, sample size, and features) | Outcome measures | Handling missing data | Algorithms | Performance | Summary |
|-------------|-----------------------------------------------------------|------------------|----------------------|------------|-------------|---------|
| Retrospective cohort (2008)<sup>46</sup> | Temporal split, development cohort (March 2005 to December 2006) and a validation cohort (January 2007 to October 2007); a cohort of hospital patients given broad-spectrum antibiotics 392 (288/104) out of 54,226 (41,224/13,002). Case:control ≈ 1:100. Reported 4 features | CDI | Not addressed | Logistic regression model to identify significant predictors individually. A scoring algorithm to create four categories of CDI risk | AUROC = 0.712 | Developed an easily implemented risk index for risk stratification of patients |
| Prospective Cohort (2013)<sup>44</sup> | Patients symptomatic of CDI/DIF between July 2014 and February 2015 from 14 Spanish hospitals 274 (Training dataset); 183 (Validation cohort). Reported 4 features | Recurrence | Not addressed | Logistic regression, Model with model calibration using Hosmer–Lemeshow test. Logistic regression to form a GEIH-CDI score | AUROC = 0.72 (0.65–0.79) | Develop a risk score for recurrent CDI prediction and stratification |
| Retrospective cohort (2019)<sup>44</sup> | First episode of adult CDI from January 1, to December 2015; For recurrence, 36 vs 191; For poor outcome, 70 vs 57 (no testing dataset) ≈35 features | Recurrence Severity | Addressed | Univariate analyses; Backward stepwise multivariate logistic regression; Multivariate regression | AUROC = 0.728/0.789 for clinical model; 0.775/0.801 for EIA-included model; 0.785/0.804 for PCR-including model for recurrence/severity | |
| Retrospective cohort (longitudinal)<sup>39</sup> (2017) | EHR (2007–2013) from single healthcare training: 9,386 incident CDI & 1,311 first CDI recurrence; testing: 1,865 incident CDI & 44 recurrent CDI 150 predictors | Recurrence (inpatient or outpatient) Right-censored data | | Univariate and bivariate regression; Competing risk discrete survival models and Cox competing risk survival regression | AUROC = 0.712 | None of the models showed a well discriminating power. Suggest including environmental and ecological predictors |
| Retrospective Cohort (2015) | Patients with lab tests positive for CFID between January 2009 and June 2013 at single healthcare system 196 CDI with 30 having CDR and breaking into 70% & 30% for training and testing 25 features | Recurrence | Not addressed | All features included Random Forest | Sensitivity (83.3%), Specificity (63.1%), and AUROC (0.826) | Expecting Random Forest model with a higher performance |
| Retrospective Cohort (2013) | January 2006 – October 2010 from 4-hospital Health Care Organization 198 out of 829 with relapse for 56 days of follow-up Reported 6 features | Relapse | Not addressed | | | |
| Retrospective Cohort (2019) | Adult inpatients diagnosed with CDI from October 2010 to January 2013 at single healthcare care: 89 out of 1144 cases of CDI having complicated CDI; 894 cases for training and 224 cases for testing 23 features for the curated model; 4271 features from EHR; final selected 900 features; 923 features for EHR + curated | Severity (3 Day Complications) | No imputation or case-wise deletion | Compared EHR-based model to one based on a small set of manually curated features; L2 regularization regression model Logistic regression | AUROC = 0.69 [0.55–0.83] on the day of CDI diagnosis; AUROC = 0.90 [0.83–0.95] 2 days after CDI diagnosis; outperformed curated feature model with AUROC = 0.84 [0.75–0.91] | Develop a model based on EHR data to accurately stratify CDI cases according to their risk of developing complications |

Continued
| Study design | Study characteristics (timeline, sample size, and features) | Outcome measures | Handling missing data | Algorithms | Performance | Summary |
|-------------|---------------------------------------------------------|------------------|----------------------|------------|-------------|---------|
| Prospective cohort \(2015\) | Discovery dataset: Boston site from December 2004 to January 2006, Validation dataset: Dublin site from November 2007 to June 2009, & Houston site from January 2006 to August 2010 for Discover and 345 for validation 3 features (Age, WBC, and Creatinine) | Severity | Not addressed | Univariate logistic regression Multivariate logistic regression analysis to form a Clostridium difficile severity score \(\text{CDSS}\) | AUROC = 0.725 \([0.675–0.769]\) | Developed a CDSS scoring system to predict severe CDI |
| Retrospective cohort \(2011\), et al. \(2011\) | January 2004 and December 2007 255 patients 4 risk factors (history of malignancy + 3 laboratory variables) | Severity | Not addressed | Univariate analysis Composite scoring: CDI severity index score | AUROC = 0.78; Sensitivity \(=82\%\); Specificity \(=65\%\) | Develop a composite score for risk stratification of severe CDI |
| Prospective cohort \(2009\) | A single healthcare 8 out of 58 for day1 and 75 for day 3 having severe complications 3 Laboratory variables | Severity | Not addressed | No feature selection Composite scoring: RUWA scoring system | Sensitivity: 80.0\% \([39.4–96.3]\) and 62.5\% \([32.3–85.6]\); Specificity: 77.4\% \([73.5–78.9]\) and 82.1\% \([78.5–84.8]\) on day 1 and day 3 respectively | the Ratio of white cell count on the day of the positive C. difficile toxin test to two days previously, as well as the Urea, White cell count and Albumin on the day of the positive C. difficile toxin test |

Table 3. Summary of the reported clinical decision tools to predict *Clostridioides difficile* infection (CDI), severity, or recurrence. We conducted a comprehensive search on PubMed and Web of Science by combining two major themes of *Clostridioides difficile* and prediction. The search strings for prediction of *Clostridioides difficile* infection were: (“clostridioides difficile” [Mesh] OR “clostridium difficile” [Mesh]) AND (“prediction” [Mesh] OR “machine learning”). We focused on human subject studies. Any review articles and studies focusing on metagenomics or microbiome data were excluded. The PubMed database was searched, and studies available between January 1, 1990, and May 31, 2021, were included. We also checked the reference of each included study, and additional studies that were missed during the initial search were appended. A total of 23 original articles were included.

The results from a subgroup with PSM suggested that the contribution of the genetic variant in model prediction was minimum in elderly patients. This finding was consistent with the decreased genetic heritability observed in late-onset compared to early-onset in multiple complex disease traits\(^23\), and rs2227306 (IL-8) was related to early-onset of disease\(^24\).

As summarized in Table 3, the majority of the predictive models for CDI\(^30–36\) are not based on large EHR or claims databases until recently\(^37–41\), whereas studies performed on recurrence\(^42–46\) or severity\(^46–50\) include very small cohorts. Further, the existing studies do not compare algorithms; instead, they focus on the amount of information extracted necessary for improved prediction. In general, the amount of information, such as the number of variables, correlates with model performance; however, including hundreds of variables can lead to models with lower interpretability and reduced generalizability to other healthcare systems. For example, some institution-specific features can rank in the top tier in feature importance; therefore, a healthcare-based model may have limited discriminative power in the prediction of individuals from other healthcare systems when geographic, social-economic, and clinical management environments differ significantly.

The generalizability of developed prediction models from a single healthcare system to others is debatable\(^37\). Since this study aims to utilize only common clinical risk factors readily available in most EHRs to build a prediction model, the conclusion made from this study could have better generalizability and may be easier to implement elsewhere. For these reasons, we propose that this integrated model is more transferable to EHR than complex models with manually curated variables and datasets.

**Strength and limitation**

The strength of this study lies in the following, (1) development of a prediction model of symptomatic CDI by integration of genetic and common clinical risk factors; (2) evaluation of several advanced ML algorithms to compare their performance; (3) identification of an association between the genetic variant and the outcome variable, which was confounded by age and sex; (4) determination of the value of the genetic feature in its contribution to the model performance, in general, and propensity score matched subgroups; and (5) identification of the selection bias in the cohort with genetic data available.

This study has some limitations, including (1) the accuracy of EHR data collection and recording processes that may vary by clinician, hospital, and over time to possibly prevent the generalizability of developed models to other healthcare systems; (2) our data came from a single healthcare system with a patient population that was predominantly European ancestry. The features selected from this homogenous population may not best represent or cover the complexity of feature space derived from a heterogeneous population; (3) we only tested a common genetic variant with a high MAF. We expect the polygenic risk score developed from the consortium-based...
GWAS with individual effect size estimated from thousands of genetic variants would better represent the genetic liability to CDI and other complex diseases.

In conclusion, we showed that developing robust prediction models for CDI, and perhaps other complex conditions, requires a step-wise approach to ensure the highest level of transparency and lowest possible systemic bias. This study leveraged CDI as a disease model to demonstrate that although genetic information may improve predictions, the benefit of including genetic feature(s) in the prediction models should be thoroughly evaluated.

Data availability
The patient-level EHR data analyzed in this study may be shared with a third party upon execution of data sharing agreement for reasonable requests. Such requests should be addressed to V. Abedi. All the codes can be found at TheDecodeLab/Prediction_of_CDI_by_EHR_and_Genetics.

Received: 10 January 2022; Accepted: 31 August 2023

Publicly online: 2 October 2023

References

1. Khanna, S. & Pardi, D. S. Clostridium difficile infection: New insights into management. Mayo Clin. Proc. 87, 1106–1117. https://doi.org/10.1016/j.mayocp.2012.07.016 (2012).

2. Berkell, M. et al. Microbiota-based markers predictive of development of Clostridioides difficile infection. Nat. Commun. 12, 2241. https://doi.org/10.1038/s41467-021-23202-0 (2021).

3. Li, J. et al. Variants at the MHC region associate with susceptibility to Clostridiodes difficile infection: A genome-wide association study using comprehensive electronic health records. Front. Immunol. 12, 638913. https://doi.org/10.3389/fimmu.2021.638913 (2021).

4. El Feghaly, R. E. et al. Markers of intestinal inflammation, not bacterial burden, correlate with clinical outcomes in Clostridium difficile infection. Clin. Infect. Dis. 56, 1713–1721. https://doi.org/10.1093/cid/cit147 (2013).

5. Del Valle, D. M. et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. Nat. Med. 26, 1636–1643. https://doi.org/10.1038/s41591-020-1031-9 (2020).

6. Jiang, Z. D. et al. A common polymorphism in the interleukin 8 gene promoter is associated with Clostridium difficile diarrhea. Am. J. Gastroenterol. 101, 1112–1116. https://doi.org/10.1111/j.1572-0241.2006.00482.x (2006).

7. Caspiere, J. et al. The presence of IL-8 +781 T/C polymorphism is associated with the parameters of severe Clostridium difficile infection. Microb. Pathog. 114, 281–285. https://doi.org/10.1016/j.micpath.2017.11.066 (2018).

8. Carey, K. W. et al. A common polymorphism in the interleukin-8 gene promoter is associated with an increased risk for recurrent Clostridium difficile infection. Clin. Infect. Dis. 51, 1406–1410. https://doi.org/10.1086/657398 (2010).

9. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMC Med. 13, 1. https://doi.org/10.1186/s12916-014-0241-x (2015).

10. Carey, D. J. et al. The Geisinger MyCode community health initiative: An electronic health record-linked biobank for precision medicine research. Genet. Med. 18, 906–913. https://doi.org/10.1038/gim.2015.187 (2016).

11. Abul-Husn, N. S. et al. Genetic identification of familial hypercholesterolemia within a single US health care system. Science 354, 6319. https://doi.org/10.1126/science.aaf7600 (2016).

12. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354, 6319. https://doi.org/10.1126/science.aaf6814 (2016).

13. Burnham, C. A. & Carroll, K. C. Diagnosis of Clostridium difficile infection: An ongoing conundrum for clinicians and for clinical laboratories. Clin. Microbiol. Rev. 26, 604–630. https://doi.org/10.1128/CMR.00016-13 (2013).

14. McDonald, L. C. et al. Clinical practice guidelines for Clostridium difficile infection in adults and children: 2017 update by the infectious diseases society of america (IDSA) and society for healthcare epidemiology of America (SHEA). Clin. Infect. Dis. 66, 987–994. https://doi.org/10.1093/cid/ciy149 (2018).

15. Yuan, W. et al. Temporal bias in case-control design: Preventing reliable predictions of the future. Nat. Commun. 12, 1107. https://doi.org/10.1038/s41467-021-21390-2 (2021).

16. Borisov, V. et al. Deep neural networks and tabular data: A survey. IEEE Trans. Neural Netw. Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3229161 (2022).

17. Zhou, Z.-H. Ensemble methods foundations and algorithms. in Ensemble Methods, 23–95 (2012).

18. Rokach, L. Introduction to ensemble learning. in Ensemble Learning Pattern Classification Using Ensemble Methods 2nd Edition, 51–104. https://doi.org/10.1142/9789811201967_0003 (2019).

19. Abedi, V. et al. Prediction of long-term stroke recurrence using machine learning models. J. Clin. Med. https://doi.org/10.3390/jcm10061286 (2021).

20. Abedi, V. et al. Predicting short and long-term mortality after acute ischemic stroke using EHR. J. Neurol. Sci. 427, 117560. https://doi.org/10.1016/j.jns.2021.117560 (2021).

21. Delong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44, 837–845 (1988).

22. Sun, X. & Xu, X. Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteric curves. IEEE Signal Process. Lett. 21, 1389–1393. https://doi.org/10.1109/LSP.2014.2337313 (2014).

23. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat. Med. 26, 549–557. https://doi.org/10.1038/s41591-020-0800-0 (2020).

24. Emonts, M. et al. Polypharmacism in genes controlling inflammation and tissue repair in rheumatoid arthritis: A case control study. BMC Med. Genet. 12, 36. https://doi.org/10.1186/1471-2350-12-36 (2011).

25. Xu, D. et al. Quantitative disease risk scores from EHR with applications to clinical risk stratification and genetic studies. NPJ Digit. Med. 4, 116. https://doi.org/10.1038/s41746-021-00488-3 (2021).

26. El Feghaly, R. E., Stauber, J. L., Tarr, P. I. & Haslam, D. B. Intestinal inflammatory biomarkers and outcome in pediatric Clostridium difficile infections. J. Pediatr. 163, 1697–1704. https://doi.org/10.1016/j.jpeds.2013.07.029 (2013).

27. Hirota, S. et al. Intrarectal instillation of Clostridium difficile toxin A triggers colonic inflammation and tissue damage: Development of a novel and efficient mouse model of Clostridium difficile toxin exposure. Infect. Immun. 80, 4474–4484. https://doi.org/10.1128/IAI.00933-12 (2012).

28. Drew, R. J. & Boyle, B. RUWA scoring system: A novel predictive tool for the identification of patients at high risk for complications from Clostridium difficile infection. J. Hosp. Infect. 71, 93–94. https://doi.org/10.1016/j.jhin.2008.09.020 (2009) [Author reply 94–95].
Acknowledgements

We would like to extend our thanks to Drs. Brodginski, Alison M; Martinez, Raquel M; Olson, Jordan; Kreis, Steve; and Heiter, Barb J for clinical data validation.

Author contributions

Study conception and design: J.L. and V.A. Data extraction and systematic literature search: J.L. D.C., V.S., V.S. Code development: J.L. Analysis of results: J.L. Data preparation and pre-processing: J.L. and V.A.; Interpretation of the findings: J.L and V.A. Preparation of the first draft of the manuscript: J.L. and V.A. Critically evaluated the manuscript: D.C., P.S., D.M.W., R.Z., and V.A. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-41753-7.

Correspondence and requests for materials should be addressed to V.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
