Journal of Community Hospital Internal Medicine Perspectives

COACTION PUBLISHING

RESEARCH ARTICLE

A prediction model for *Clostridium difficile* recurrence

Francis D. LaBarbera, MD*†, Ivan Nikiforov, MD, Arvin Parvathenani, MD, MS, Varsha Pramil, BS and Subhash Gorrepati, MD, MS

Department of Internal Medicine, PinnacleHealth Hospital, Harrisburg, PA, USA

Background: *Clostridium difficile* infection (CDI) is a growing problem in the community and hospital setting. Its incidence has been on the rise over the past two decades, and it is quickly becoming a major concern for the health care system. High rate of recurrence is one of the major hurdles in the successful treatment of *C. difficile* infection. There have been few studies that have looked at patterns of recurrence. The studies currently available have shown a number of risk factors associated with *C. difficile* recurrence (CDR); however, there is little consensus on the impact of most of the identified risk factors.

Methods: Our study was a retrospective chart review of 198 patients diagnosed with CDI via Polymerase Chain Reaction (PCR) from January 2009 to Jun 2013. In our study, we decided to use a machine learning algorithm called the Random Forest (RF) to analyze all of the factors proposed to be associated with CDR. This model is capable of making predictions based on a large number of variables, and has outperformed numerous other models and statistical methods.

Results: We came up with a model that was able to accurately predict the CDR with a sensitivity of 83.3%, specificity of 63.1%, and area under curve of 82.6%. Like other similar studies that have used the RF model, we also had very impressive results.

Conclusions: We hope that in the future, machine learning algorithms, such as the RF, will see a wider application.

Keywords: *Random Forest*; *machine learning algorithm*; *hospital infection*

*Correspondence to: Francis D. LaBarbera, PinnacleHealth Hospital, Department of Internal Medicine, 205 S. Front St. 3rd floor, Harrisburg, PA 17110, USA, Email: flabarbera@pinnaclehealth.org

Received: 16 September 2014; Revised: 3 December 2014; Accepted: 8 December 2014; Published: 3 February 2015

Clostridium difficile is a spore-forming, gram-positive, bacillus bacterium that is associated with severe and often life-threatening infections and inflammation of the colon. The disease processes can range from mild diarrhea to fulminant colitis and death. *C. difficile* has become the most frequent cause of nosocomial diarrhea in the United States. According to the Centers for Disease Control, the infection is responsible for over 14,000 deaths per year. Because of more virulent strains and evolving antimicrobial resistance, the rates of incidence and recurrence have been increasing (1). In addition, *C. difficile* infection (CDI) is a heavy burden on health care expenses and accounts for an increased use of medical resources. Recent studies have shown that health care costs ranged from $2,871 to $4,846 per case for primary CDI and from $13,655 to $18,067 per case for recurrent CDI (2).

Factors, which are known to alter the normal enteric flora, are associated with risk of *C. difficile* colonization (3). Although the predominant risk factor among them is associated with antibiotic therapy (4), other postulated risk factors such as, advanced age, chronic illnesses or comorbidities, hospitalizations, non-surgical gastrointestinal procedures, chemotherapy, and other immunosuppressants, play a major role in altering the flora and subsequent acquisition of CDI (5). There have been multiple publications that have demonstrated the association of these different variables with the acquisition of CDI and subsequent re-infections; however, the use of an organized machine learning, sensitivity analysis approach, such as a Random Forest (RF) statistical model, has not been used. In our study, we emulated the techniques of Amalakuhan et al. (6) and the prediction model they created using an RF model in predicting patients at risk for chronic

†Francis D. LaBarbera is the guarantor for all parts of this paper.

Citation: Journal of Community Hospital Internal Medicine Perspectives 2015, **5**: 26033 - http://dx.doi.org/10.3402/jchimp.v5.26033

obstructive pulmonary disease (COPD) exacerbation. We employed the RF machine learning algorithm to predict *C. difficile* recurrence (CDR).

Methodology

Definitions

'Recurrence of *C. difficile*' was defined as confirmed presence of *C. difficile* toxin via polymerase chain reaction (PCR) after complete resolution of diarrhea for a minimum of 2 weeks and a maximum of 6 months and the completion of antibiotic therapy (5, 7).

RF is a statistical model used for classification and regression which works by using input variables to create an output of regression trees. The mode class of all of the classes of the individual trees is then chosen as the output and is used to determine the variables of interest.

Regression trees are binary trees with nodes that correspond to different values in the input variables. These trees are developed using a training set. At each node or branch, the RF algorithm searches for a value that best separates all instances within that node based on the outcome of interest. If the instance chosen is not able to be separated further, then it is called a terminal node. This process is repeated until all instances are terminal nodes (6).

Variables

The variables used in this study were selected after an extensive review of literature and only the most common comorbidities found in patients with CDI and re-infection were included. Among the numerous existing variables, 25 of the most strongly associated variables have been selected. Variables which were not found to be significant have not been included in order to optimize the RF algorithm. Table 1 depicts the significant variables used in the RF model in this study.

Table 1. Explanatory variables used in this study

Age	Smoking
Coronary artery disease	Low-risk antibiotics
Chronic kidney disease	High-risk antibiotics
Gastrointestinal (GI) malignancy	H_2 antagonist
Gender	Alcohol use
Peptic ulcer disease	No GI surgery
Inflammatory bowel disease	One GI surgery
Immunosuppression	Two GI surgeries
Race	Hypertension
Gastro esophageal reflux disease	Proton pump inhibitor (PPI) 20 mg
Corticosteroids	PPI 40 mg
Chemotherapy	PPI 80 mg
Diabetes	

Sample selection

A retrospective chart review was performed on patients diagnosed with CDI based on International Classification of Disease (ICD-9) codes. The selection of the study population, selection criteria, and sampling were all performed subject to the approval of the institutional review board (IRB).

Inclusion criteria

Patients diagnosed with CDI via PCR between January 2009 and June 2013.

Exclusion criteria

1. Patients with recurrence within 2 weeks or after 6 months of initial CDI
2. Patients with documented 'non-compliance' to prescribed medical therapy

Sample size

Using these criteria, data of 200 randomized patients diagnosed with CDI were collected. Two patients were subsequently removed due to our set exclusion criteria. The prevalence of CDR within the randomly selected sample size was 15% (30 patients).

Data analysis

RF statistical analysis and randomization was performed using the SPM Salford Predictive Modeler® version 6.0 (Salford System, San Diego, CA). The predictive model was designed by professors Leo Breiman and Adele Cutler of the University of California, Los Angeles. Our sample population was randomly separated into two distinct groups, a training group and a validation group. The training group comprised 70% of our patients. The remaining 30% were placed in the validation group. The training group was used essentially to create our 'learning algorithm', which comprised 2,000 regression trees. Each tree comprised binary nodes or branches, which contribute their results to the variables of interest. At each node, a variable is tested (e.g., 80 mg PPI). At that node, the data entered will either fall into a branch of CDR or no recurrence. Through the 2,000 regression trees, the outcome of interest (CDR) will be distinguished. The cumulative predictions created in the 2,000 trees will create the probability of the patient having a recurrence of an initial CDI.

The training and testing groups, although broken into two groups, still allow for all patients in the sample selection to be run through the predictive model. The 30% represents only one validation group in one of the 500 runs. This group essentially can be completely different from one run to the next. This statement remains the same for the 70% training group. As opposed to Amalakuhan's study, which used a 75% training group to 25% validation group distribution split (8, 9), we used an experimental 70%:30% distributional split. This allowed an additional 10 patients

Citation: Journal of Community Hospital Internal Medicine Perspectives 2015, **5**: 26033 - http://dx.doi.org/10.3402/jchimp.v5.26033

to our validation group, presumptively strengthening the end result and the predictive probability of our model.

Within both the training and validation groups, the predicted probability of determining the patients who had CDR was almost completely equal. The RF model using the indices as set above was run 500 times, and for each of these 500 runs, the accuracy of the predictive model was assessed by calculating the sensitivity, specificity, area under the curve (AUC) for receiver operating characteristic, and precision (Fig. 1).

Results

Our population consisted of 198 patients from two sister hospitals in a community setting with a documented CDI. Of those patients, 30 had CDR, giving us a recurrence rate of approximately 15%. The mean age of the patients in the recurrence group was 70.8 (SD 17), while the mean age of the population without recurrence was 68.7 (SD 17.7). The majority of the population was Caucasian (88.4%) with the remainder of the population being either African American (6.6%) or other races (5.0%) (Table 2).

Our final model based on 500 runs of the RF produced the following results: sensitivity, 83.3%; specificity, 63.1%; overall correct predicted percentage, 66.1%; and AUC, 82.6%

Discussion

CDI is a growing concern in the health care system, and it represents one of the most difficult challenges faced by clinicians. Throughout the 1990s and 2000s, the number of *C. difficile*-related hospital stays per 100,000 population has been steadily increasing, peaking at 114.6 in 2008 from a low of 33.2 in 1993 (10). The major problem associated with *C. difficile* infection is the high rate of recurrence of 20 to 35%. If a recurrence has occurred, the chances of a second recurrence further increases to 45–65% (11, 12). CDR increases mortality and morbidity, length of hospital stay, health care costs, and utilization of other health care resources. It also puts additional burden on the patient's quality of life and their families. Previous studies have also documented that CDR resulted in 265 additional days of vancomycin use and 19.7 days of metronidazole use (13). Although some studies have looked at the causes, there is little consensus on causes of CDR. Previous studies have identified the following associated risk factors: age, Horn's index, proton pump inhibitor use, antibiotic use, alteration of colonic microflora, initial disease severity, and hospital exposure (14–18).

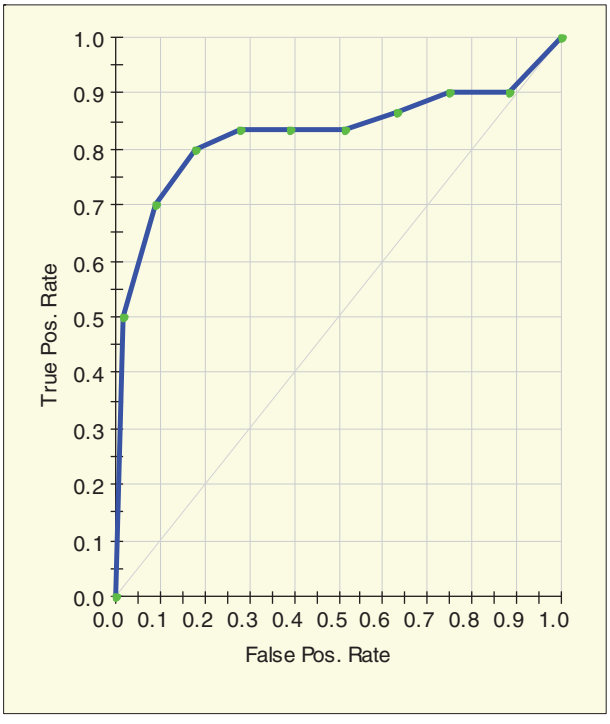

Fig. 1. Receiver operating curve with area under curve.

For this study, we created an RF predictive model using multiple well-known risk factors associated with CDI identified in previous studies. After 500 runs of the tree-based algorithm, the RF model performed extremely well in classifying predictors of CDR versus non-recurrent cases. The overall accuracy of our RF model was 66.1% with a sensitivity and specificity of 83.3 and 63.1%, respectively. The area under the receiver operating curve was 82.6%, which is comparable to other strong models. Our literature review did not reveal many studies using prediction models to identify CDR. Hu et al. (15) showed similar results. Our study expanded on their work by using a larger amount of patients that were followed for a longer period of time of 3 years. In addition, specific important co-variables such as PPI's chemotherapy, corticosteroids, and antibiotic use were used in our study (19). We also had a substantially higher sensitivity and AUC.

The RF has already been used successfully in other studies in determining predictors in various chronic and acute conditions. The study done by Amalakuhan et al. (6) previously demonstrated that the RF was excellent at predicting factors associated with readmissions for COPD exacerbation. A well-documented study by Adrienne Chu demonstrated that the RF machine learning algorithm outperformed six other prediction models in determining the cause of gastrointestinal bleeding (20). The models that were outperformed included well-known algorithms such as the support vector machine, boosting, artificial neural network, linear discriminant analysis, and logistic regression.

One of the major strengths of our study was that it contained a large number of patients who were followed for several years (2009–2013). We also used a large number of significant variables in this study to create the prediction model. The study had a high AUC and high sensitivity.

Table 2. Demographic and risk factor table

	No recurrence		Recurrence		*p*
Total number of patients	168		30		
Demographics					
Age (mean, SD, range)	68.7 (17.7)	3–96	70.8 (17.0)	28–93	0.5339
Male (number,%)	62	36.69	17	56.67	0.0393
Race					
Caucasian	151	89.35%	25	83.33%	0.3544
African American	9	5.33%	4	13.33%	0.1126
Others	9	5.33%	1	3.33%	1
Risk factors					
Smoking	60	35.50%	8	26.67%	0.3364
ETOH	43	25.44%	8	26.67%	0.8876
Hypertension	133	78.70%	22	73.33%	0.5141
CAD	78	46.15%	12	40.00%	0.5326
CKD	51	30.18%	11	36.67%	0.4794
Diabetes	56	33.14%	12	40.00%	0.4651
GERD	46	27.22%	8	26.67%	0.9500
PUD	11	6.51%	3	10.00%	0.4479
IBD	3	1.78%	0	0.00%	1.0000
IBS	6	3.55%	1	3.33%	1.0000
GI cancer	5	2.96%	1	3.33%	1.0000
Immunosuppressed	23	13.61%	4	13.33%	1.0000
Low-risk antibiotics	55	32.54%	11	36.67%	0.6585
High-risk antibiotics	104	61.54%	17	56.67%	0.6145
H_2 antagonist	13	7.69%	1	3.33%	0.6988
Corticosteroids	26	15.38%	9	30.00%	0.0527
Chemotherapy	14	8.28%	1	3.33%	0.7022
Number of GI surgeries					
One	35	20.71%	12	40.00%	0.0219
Two or more	16	9.47%	3	10.00%	1.0000
None	118	69.82%	15	50.00%	0.0336
PPI dose					
20 mg	16	9.47%	4	13.33%	0.5126
40 mg	117	69.23%	16	53.33%	0.0883
80 mg	1	0.59%	3	10.00%	0.0113

One of the major limitations of our study was that the majority of our sample population was Caucasian.

Conclusion

C. difficile is a growing problem in the hospital and community setting. The recurrent nature of infection is worrisome since repeated use of antibiotics against the same strain of bacteria may lead to resistance mechanisms. In this study, we used the RF machine learning algorithm to create a strong prediction model with high sensitivity to predict CDR. In the evolving field of medical informatics, the use of such learning algorithm models can be used in risk factor stratification for hospitalized patients. If patients at risk for CDR could be accurately identified, specific management strategies could be developed resulting in better management, decreased morbidity and mortality, better health care resource utilization, and decreased length of hospital stay. We believe that the advantages offered by the RF makes it the ideal tool for this task.

Acknowledgements

We thank Yijin Wert, MS, biostatistician for PinnacleHealth Hospital Systems. We also thank Franklin Fontem, MD, and Anix Vyas, MD, for their assistance in editing and data gathering for this project.

Conflicts of interest and funding

The authors report that there are no conflicts of interest in publishing this paper. No private funding was received for this research project.

References

1. Freeman J, Bauer MP, Baines SD, Corver J, Fawley WN, Goorhuis B, et al. The changing epidemiology of *Clostridium difficile* infections. Clin Microbiol Rev 2010; 23(3): 529.
2. Ghantoji SS, Sail K, Lairson DR, DuPont HL, Garey KW. Economic healthcare costs of *Clostridium difficile* infection: A systematic review. J Hosp Infect 2010; 74(4): 309–18.
3. Cohen SH, Gerding DN, Johnson S, Kelly CP, Loo VG, McDonald LC, et al. Clinical practice guidelines for *Clostridium difficile* infection in adults: 2010 update by the Society for Healthcare Epidemiology of America (SHEA) and the infectious diseases society of America (IDSA). Infect Contr Hosp Epidemiol 2010; 31(5): 435–55.
4. McFarland LV. Alternative treatments for *Clostridium difficile* disease: What really works? J Med Microbiol 2005; 54(Pt 2): 101–11.
5. Surawicz CM, McFarland LV, Greenberg RN, Rubin M, Fekety R, Mulligan ME, et al. The search for a better treatment for recurrent *Clostridium difficile* disease: Use of high-dose vancomycin combined with *Saccharomyces boulardii*. Clin Infect Dis 2000; 31(4): 1012–17.
6. Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriyath P, Fischman D. A prediction model for COPD readmissions: Catching up, catching our breath, and improving a national problem. J Community Hosp Intern Med Perspect 2012; 2: 9915.
7. Pépin J, Routhier S, Gagnon S, Brazeau I. Management and outcomes of a first recurrence of *Clostridium difficile*-associated disease in Quebec, Canada. Clin Infect Dis 2006; 42(6): 758–64.
8. R Project. R Project contributors. Available from: http://www.r-project.org/ [cited 27 February 2011].
9. Random Forest package for R project. Fortran original: Leo Breiman, Adele Cutler, R port: Andy Liaw, Matthew Wiener. Available from: http://cran.r-project.org/web/packages/randomForest/index.html [cited 27 February 2011].
10. Lucado J, Gould C, Elixhauser A. *Clostridium difficile* infections (CDI) in hospital stays, 2009. HCUP Statistical Brief #124. Rockville, MD: Agency for Healthcare Research and Quality; Available from: http://www.hcup-us.ahrq.gov/reports/statbriefs/sb124.pdf [cited January 2012].
11. McFarland LV, Elmer GW, Surawicz CM. Breaking the cycle: Treatment strategies for 163 cases of recurrent *Clostridium difficile* disease. Am J Gastroenterol 2002; 97(7): 1769–75.
12. Zilberberg MD, Reske K, Olsen M, Yan Y, Dubberke ER. Risk factors for recurrent *Clostridium difficile* infection (CDI) hospitalization among hospitalized patients with an initial CDI episode: A retrospective cohort study. BMC Infect Dis 2014; 14: 306.
13. Rohlke F, Stollman N. Fecal microbiota transplantation in relapsing *Clostridium difficile* infection. Therap Adv Gastroenterol 2012; 5(6): 403–20.
14. Eyre DW, Walker AS, Wyllie D, Dingle KE, Griffiths D, Finney J, et al. Predictors of first recurrence of *Clostridium difficile* infection: Implications for initial management. Clin Infect Dis 2012; 55(Suppl 2): S77–87.
15. Hu MY, Katchar K, Kyne L, Maroo S, Tummala S, Dreisbach V, et al. Prospective derivation and validation of a clinical prediction rule for recurrent *Clostiridium difficle* infection. Gastroenterology 2009; 136: 1206–14.
16. Kelly JP. Can we identify patients at high risk of recurrent *Clostridium difficile* infection? Clin Microbiol Infect 2012; 18(Suppl 6): 21–7.
17. Fekety R, McFarland LV, Surawicz CM, Greenberg RN, Elmer GW, Mulligan ME. Recurrent *Clostridium difficile* diarrhea: Characteristics of and the risk factors for patients enrolled in a prospective, randomized, double-blinded trial. Clin Infect Dis 1997; 24(3): 324–33.
18. Samie AA, Traub M, Bachmann K, Kopischke K, Theilmann L. Risk factors for recurrence of *Clostridium difficile*-associated diarrhea. Hepatogastroenterology 2013; 60(126): 1351–4.
19. Freedberg DE, Salmasian H, Friedman C, Abrams JA. Proton pump inhibitors and risk for recurrent *Clostridium difficile* infection among inpatients. Am J Gastroenterol 2013; 108: 1794–801.
20. Chu A, Ahn H, Halwan B, Artifon ELA, Barkun A, Lagoudakis MG, et al. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. J Artif Intell Med 2008; 42(3): 247–59.