Am J Digest Dis 2023;9(2):12-21
www.ajdd.us /ISSN:2329-6992/AJDD0150901

Original Article

Association of enteric infections and disease activity in inflammatory bowel disease: a retrospective study utilizing machine learning techniques

Christopher Kim[1], Vamsi Maturi[3], Adam Saleh[3], Manuel Garza[2], Medha Narwankar[3], Caroline Perry[1], Kerri Glassner[1], Bincy Abraham[1]

[1]Division of Gastroenterology, Department of Medicine, Houston Methodist Hospital, Houston, TX, USA; [2]Department of Medicine, Houston Methodist Hospital, Houston, TX, USA; [3]Texas A&M Health Science Center, College Station, Houston, TX, USA

Received April 24, 2023; Accepted November 16, 2023; Epub December 15, 2023; Published December 30, 2023

Abstract: Enteric infections are frequently encountered in patients with active IBD symptoms. Few studies have evaluated endoscopic findings in symptomatic IBD patients with an enteric infection. We hypothesized that IBD patients with an enteric infection were more likely to have active inflammation on colonoscopy. Machine learning techniques were used to predict the presence of enteric infection in IBD patients with active disease on colonoscopy. Patients with IBD seen from 2015 to 2020 at Houston Methodist Inflammatory Bowel Disease program were identified. Those who had stool PCR testing performed for evaluation of diarrhea were included in the study. Retrospective data collection included demographic data, disease subtype, disease location, laboratory data, clinical and endoscopic findings. Machine learning techniques were used to help identify predictors of the presence of an enteric infection. There were 284 patients with at least one stool PCR test and among these, 167 (58.8%) patients had an infection identified. Those with ulcerative colitis (UC), particularly pancolitis, were more likely to have an infection than those with Crohn's disease (CD). Both UC and CD patients with inflammation identified on colonoscopy (based on endoscopic score) were more likely to have an enteric infection. Finally, a multivariate analysis using machine learning techniques showed that age predicted likelihood of enteric infection in IBD patients. Enteric infections were commonly identified in IBD patients being evaluated for diarrhea. *Clostridioides difficile and Escherichia coli* species were most common. UC, particularly pancolitis, and endoscopic disease severity increased the likelihood of enteric infection. Age was also a significant predictor as shown in our multivariate analysis. Further tailoring of machine learning techniques with larger patient numbers and additional variables are future research areas of interest.

Keywords: Clostridioides difficile infection, Crohn's disease, enteric infection, inflammatory bowel disease, ulcerative colitis, machine learning

Introduction

Inflammatory bowel disease (IBD) encompasses two chronic inflammatory diseases: Crohn's disease (CD) and ulcerative colitis (UC). The prevalence of both CD and UC is on the rise in the United States, with over 0.7% of the population now affected [1-3]. Although the exact etiology of IBD remains unclear, several studies have shown that enteric infections, the innate immune system, and the intestinal microbiome are associated with the pathogenesis and exacerbation of IBD [4-8]. Diagnosis of UC or CD rely on a combination of presentation and tests such as endoscopy, radiology and histology [9, 10]. Colonoscopy, with terminal ileum intubation and histologic analysis, remains the primary diagnostic method for both diseases. Laboratory testing for IBD includes measuring markers of inflammation such as fecal calprotectin or lactoferrin, C reactive protein (CRP), and erythrocyte sedimentation rate (ESR). These laboratory tests are commonly used to monitor disease activity and progression. The management of IBD involves achieving clinical and endoscopic remission with evolving goals of reaching higher targets such as histologic remission in UC and transmural healing in CD.

Ideal therapeutic strategy would reflect multiple factors including disease location, disease activity, patient preferences, and comorbidities.

Among other potential environmental contributions to the pathogenesis of IBD, enteric infections are thought to contribute to the generation of chronic inflammation by triggering an already altered immune response [9-11]. Specific pathogens connected to IBD in prior studies include *Salmonella* species, *Clostridioides difficile*, enteropathogenic *Eschericia coli* (EPEC), *Listeria monocytogenes*, *Campylobacter species*, and *Mycobacteria* species, as well as other parasites and viruses [12-14]. One study found acute gastroenteritis leads to a 2.4 folds increased risk of IBD (95% CI 1.7-3.3), with the greatest risk observed in the first year of follow-up [9]. Both identifying microorganisms and quantifying levels of inflammation are routinely employed in caring for patients with IBD who are having diarrhea [15]. Molecular techniques, specifically stool polymerase chain reaction (PCR) testing in conjunction with clinical reasoning, have improved the diagnosis of enteric infections [16, 17].

Even with these advancements, enteric infections and IBD exacerbations can be difficult to distinguish given the similarities in presenting symptoms and laboratory results. This confound can hinder treatment choices. Machine learning techniques have been used previously to predict the presence of IBD based on stool and serologic testing. ML models have also been used to help identify *Clostridioides difficile* (*C difficile*) infection in hospitalized patients [18]. Specifically, gradient-boosted decision trees and neural networks were able to predict *C difficile* infections with high levels of discrimination [19]. Machine learning techniques are positioned to elucidate key relationships in patient characteristics and disease presence. However, building and implementing machine learning requires data. Potential data sources include electronic health records in the form of clinical scores, characteristics, and laboratory reports, imaging results. Another source of data is high throughput datasets such as genomics, epigenomics, and proteomics. These data sources can be used to implement machine learning models aimed at differentiating disease [18]. Currently, implementation of machine learning in IBD is limited to research and proofs-of-concept. For example, Yuan et al. was able to identify candidate genes related to IBD using gene expression profiles of 85 IBD patients and 42 healthy controls in a support vector machine learning algorithm [20]. There are similar examples on the use of machine learning in IBD, but to our knowledge this is the first to explore machine learning in predicting enteric infections in IBD cohorts.

The primary aim of this study was to evaluate the association between enteric infections and IBD disease activity. The secondary aim of this study was to utilize machine learning techniques as a proof of concept to predict the presence of enteric infections based on individual IBD phenotypes.

Patients and methods

Study population

We retrospectively identified a cohort of 574 unique IBD patients seen between 2015 and 2020 at Houston Methodist Fondren IBD Program located in the Texas Medical Center. Patients with a biopsy-proven diagnosis of UC, CD or indeterminate colitis, who had a stool PCR test performed for diarrhea and were evaluated endoscopically within three months of the stool PCR test were included in the study. Patients without endoscopic findings within three months were excluded. Endoscopic scoring and clinical characteristics among patients with positive and negative stool PCR testing for infection were compared. All patients with an infection identified on stool PCR received antimicrobial therapy unless it was contraindicated due to the type of infection.

Data collection

A total of 284 out of 574 unique IBD patients were identified in both the inpatient and outpatient setting who had at least one stool PCR ordered. Specifically, patients with symptoms of diarrhea or increased urgency or frequency of bowel movements underwent testing with stool PCR through a BioFire FilmArray™ gastrointestinal panel. The data includes patient characteristics, including age, sex, disease subtype, disease location based on the Montreal Classification, medications, and endoscopic findings within three months of stool

testing. We recorded the specific infection identified on positive stool PCR testing and analyzed stool lactoferrin values if collected within three months of the stool PCR test. The evaluation of endoscopic activity incorporated the Mayo endoscopic score for ulcerative colitis patients, graded on a scale of 0 to 3. The scoring was as follows: 0 for remission or absence of visible inflammation; 1 for mild inflammation, evident through erythema, decreased vascular pattern, or mild friability; 2 for moderate inflammation characterized by marked erythema, absent vascular pattern, friability, or the presence of erosions; and 3 for severe inflammation, identified by spontaneous bleeding or ulceration. For patients with Crohn's disease post ileo-cecal resection, the Rutgeerts' score was utilized. The grading was as follows: i0 for no recurrence of inflammation; i1 for 5 or fewer aphthous ulcers; i2 for more than 5 aphthous ulcers with skip areas of larger lesions, or lesions confined to the ileocolonic anastomosis; i3 for diffuse aphthous ileitis with diffusely inflamed mucosa; and i4 for diffuse inflammation with larger ulcers, nodules, or narrowing. For the remaining patients with Crohn's, the Simple Endoscopic Score for Crohn's Disease (SES-CD) was utilized, with scoring ranging from 0-56 based on the absence or presence of ulcers, the percentage of ulcerated surface, the percentage of affected surface, and narrowing by segment including the ileum, right colon, transverse colon, left colon, and rectum. These scores were converted to an ordinal scale of disease activity levels: remission, mild, moderate, and severe. Specifically, Mayo endoscopic scores were categorized as 0 for remission, 1 for mild, 2 for moderate, and 3 for severe disease activity. Rutgeerts' scores were interpreted as i0-i1 for remission, i2 for moderate, and i3-i4 for severe disease activity. SES-CD scores were translated as 0-2 for remission, 3-6 for mild disease, 7-15 for moderate disease, and scores above 15 indicating severe disease activity.

Endoscopic scoring was converted from Mayo endoscopic score (used for ulcerative colitis and based on 0 for remission or no visual inflammation, 1 for mild inflammation described as erythema, decreased vascular pattern and/or mild friability, 2 for moderate inflammation described as marked erythema, absent vascular pattern, friability and/or erosions, and 3 for severe inflammation defined by spontaneous bleeding and/or ulceration), Rutgeert's score (used in Crohn's disease after an ileo-cecal resection and graded as i0 for no recurrence of inflammation, i1 for 5 or fewer aphthous ulcers, i2 for more than 5 aphthous ulcers, skip areas of larger lesions or lesions confined to the ileocolonic anastomosis, i3 for diffuse aphthous ileitis with diffusely inflamed mucosa, i4 for diffuse inflammation with larger ulcers, nodules and/or narrowing), or Simple Endoscopic Score for Crohn's Disease (SES-CD) (based on the absence or presence of ulcers, the percentage of ulcerated surface, percentage of affected surface, and narrowing by segment including ileum, right colon, transverse colon, left colon and rectum with a total score of 0-56) to an ordinal scale with the following disease activity levels: remission, mild, moderate and severe. Mayo endoscopic scores from 0-3 were converted as follows: 0 = remission, 1 = mild, 2 = moderate, 3 = severe disease activity. Rutgeert's scores from i0-i4 were converted as i0-i1 = remission, i2 = moderate, i3-i4 = severe. SES-CD scores were converted as 0-2 = remission, 3-6 = mild disease, 7-15 = moderate disease, > 15 = severe disease activity.

Statistical analysis

Descriptive statistics - presented as absolute numbers, percentages, and standard deviations for categorical variables - were used to summarize the characteristics of patients included in the dataset. A data scientist manually reviewed and imported the data to a Python pandas data frame for data mapping, transformation, and manipulation. We analyzed each patient subset, stratified by positive or negative PCR test, via the Student's t-test and each categorical subset via X^2 (Chi-squared) analysis. *P*-values for the respective tests were calculated with the SciPy stats package. Statistical significance required p-values < 0.05.

Multivariate analysis

To develop predictive models of patients with enteric infections, we conducted a multivariate analysis with machine learning techniques to evaluate our data. We formulated the analysis as a binary classification problem with a balanced random forest classifier. We selected a tree-based model due to its effectiveness in relatively small datasets when compared to

Table 1. Patient characteristics

Characteristics	Number (%) (N = 284 patients)
Sex - no. (%)	
Male	175 (61.6%)
Female	109 (38.4%)
Age - YR	
Median	41
Range	18-85
Disease Subtype - no. (%)	
Crohn's	166 (58.5%)
Ulcerative Colitis	118 (41.5%)

other machine learning models such as neural networks [21].

To ensure that IBD disease activity was included in the analysis, only patients with an endoscopy score within the last 3 months were included. Patient age, sex, disease subtype, disease location, lactoferrin, and endoscopy results were used as inputs to our classifier from the Scikit-learn package in Python 3.6. Variables with ≥ 40% null values across the data set were excluded from the analysis. We hot encoded categorical variables with the get_dummies function from the Scikit-learn package and then removed collinear variables prior to model training and evaluation.

The target variable was stool PCR positivity with K-fold cross validation used to evaluate the model's performance with mean f1 scores, mean recall, mean precision, and a receiver operating characteristic (ROC) curve. Additionally, Gini importances were reported for included variables. Hyperparameters of our balanced random forest classifier were then optimized to maximize model performance.

Ethical considerations

This study was conducted in compliance with the Health Insurance Portability and Accountability Act, and Institutional Review Board (IRB) approval was obtained, as per institutional policy.

Results

Patient characteristics

574 unique IBD patients were seen at our clinic with at least one stool panel collected. 284 of these patients with at least one completed GI molecular pathogen PCR panel were randomly chosen to be included in our study. The median age of patients in the cohort was 41 years old; 175 (61.6%) were male, and 109 (38.4%) were female. In our cohort, 166 (58.5%) patients had Crohn's disease, and 118 (41.5%) had ulcerative colitis (**Table 1**).

Table 2. Enteric infections detected by stool PCR

PCR result	Number (%) (N = 284)
Patients with positive Stool PCR	167 (58.8%)
Patients with negative Stool PCR	117 (41.2%)
Clostridioides difficile	105 (62.8%)
Escherichia coli	34 (20.4%)
Enteroaggregative E. coli	5 (3.0%)
Enteropathogenic E. coli	27 (16.2%)
Enterotoxigenic E. coli	2 (1.2%)
Plesiomonas shigelloides	2 (1.2%)
Campylobacter	5 (3.0%)
Vibrio	1 (0.6%)
Giardia intestinalis (lamblia)	1 (0.6%)
Multiple	5 (3.0%)
Sapovirus	3 (1.8%)
Yersinia Enterocolitica	2 (1.2%)
Astrovirus	1 (0.6%)
Cyclospora	3 (1.8%)
Norovirus	5 (3.0%)

Positive stool pathogen PCR

In total, 167/284 (58.8%) patients had a stool PCR test that was positive for an enteric infection (**Table 2**). Of those, 105 (62.8%) PCR tests were positive for *Clostridioides difficile*. Among the 62 (37.2%) other positive PCR tests, 34 (20.4%) were positive for *Escherichia coli*. *Enteropathogenic E. coli* was the subset of *E. coli* detected most often with 27 (16.2%) of positive stool PCRs, followed by *Enteroaggregative E. coli* with five (3.0%) and *Enterotoxigenic E. coli* with two (1.2%). The other most common pathogens detected included *Campylobacter* in five (3%), *Norovirus* in five (3%), *Cyclospora* in three (1.8%) and *Sapovirus* in three (1.8%) patients.

Enteric infections and associated variables

The average age of patients with a positive stool PCR was 44.3 (± 15.3) years old. Patients aged 20-50 were more likely than other age

Table 3. Age distribution amongst patients with collected stool PCR tests

Age	Positive Stool PCR (N = 167)	Negative Stool PCR (N = 117)	Chi-Square *P*-Value
[10, 20]	5 (3.0%)	7 (6.0%)	0.0034
[20, 30]	41 (24.6%)	19 (16.2%)	
[30, 40]	43 (25.7%)	18 (15.3%)	
[40, 50]	32 (19.2%)	15 (12.8%)	
[50, 60]	27 (16.2%)	34 (29.1%)	
[60, 70]	17 (10.2%)	16 (13.7%)	
[70, 80]	2 (12.0%)	7 (6.0%)	
[80, 100]	0 (0%)	1 (0.9%)	

Table 4. IBD disease subtype in patients with and without enteric infections as identified by stool PCR testing

Disease Subtype	Positive Stool PCR (N = 167)	Negative Stool PCR (N = 117)	Odds Ratio	Fisher's Exact Test *P*-Value
Crohn's Disease	86 (51.8%)	80 (48.2%)	0.492	0.00496
Ulcerative Colitis	81 (68.6%)	37 (31.4%)	2.031	

IBD-inflammatory bowel disease; PCR-polymerase chain reaction.

groups to have a positive stool PCR (P = 0.034) (**Table 3**). In patients with ulcerative colitis, 81 (68.6%) had a positive stool PCR and 37 (31.4%) were negative for a stool PCR; comparatively, in patients with Crohn's disease, 86 (51.8%) had a positive stool PCR and 80 (48.2%) had a negative stool PCR. Based on these results, patients with a positive stool PCR were significantly more likely to have ulcerative colitis than Crohn's disease (OR = 2.031 [CI 1.209-3.45], P = 0.00496) (**Table 4**).

Endoscopic findings and stool PCR positivity

Endoscopic findings were reviewed within three months of both positive and negative GI pathogen PCR tests. Findings were analyzed first by overall disease severity of IBD, classified as remission, mild, moderate or severe. Patients with severe disease endoscopically were more likely to have a positive PCR stool test, as 32 (68.1%) patients had a positive stool PCR, while 15 (31.9%) had a negative PCR test (OR = 2.13). Among patients with mild disease, 20 (55.6%) had a positive stool PCR test, and 16 (44.4%) had a negative stool PCR (OR = 1.25). Among patients with moderate disease, 10 (43.5%) had a positive stool PCR, and 13 (56.5%) had a negative stool PCR (OR = 0.77). Chi-square analysis showed a statistically significant association between disease severity and PCR positivity in evaluated patients with IBD (P = 0.0298) (**Table 5**).

Separate subset analyses evaluating patients with Crohn's disease (**Table 5**) and ulcerative colitis (**Table 5**) also showed that patients with severe disease were more likely to have a positive stool PCR test, regardless of the subtype of IBD. Among patients with severe Crohn's disease, 15 (65.2%) had positive stool PCRs, while eight (34.7%) were negative (OR = 1.875). In patients with mild endoscopic disease, 15 (60.0%) patients had a positive stool PCR, compared to 10 (40.0%) with a negative stool PCR (OR = 1.5), and in patients with moderate disease, five (41.7%) had a positive stool PCR, and seven (58.3%) patients had a negative stool PCR (OR = 0.71). Chi-square analysis showed a statistically significant association between disease severity and PCR positivity in evaluated patients with Crohn's disease (P = 0.0241).

In patients with ulcerative colitis in the subset analysis, those with severe endoscopic disease were more likely to have a positive stool PCR: 17 (70.8%) patients had a positive stool PCR, while seven (29.2%) had a negative stool PCR (OR = 2.43). Among patients with endoscopic remission, seven (70.0%) patients had a positive stool PCR, and three (30.0%) had a negative stool PCR (OR = 2.33). Patient groups with mild and moderate disease each had five (45.4%) positive and six (54.6%) negative stool PCRs (OR = 0.83). Chi-square analysis showed that the association between disease severity

Table 5. Endoscopic findings in all IBD, Crohn's, and Ulcerative Colitis patients with and without enteric infections as identified by stool PCR

Disease	Disease Severity	Positive Stool PCR	Negative Stool PCR	Odds Ratio	Chi-Square *P*-Value
All IBD (N = 148)	Remission	16 (38.1%)	26 (62.0%)	0.62	0.0298
	Mild	20 (55.6%)	16 (44.4%)	1.25	
	Moderate	10 (43.5%)	13 (56.5%)	0.77	
	Severe	32 (68.1%)	15 (31.9%)	2.13	
Crohn's Disease (N = 92)	Remission	9 (28.1%)	23 (71.9%)	0.39	0.0241
	Mild	15 (60.0%)	10 (40.0%)	1.5	
	Moderate	5 (41.7%)	7 (58.3%)	0.71	
	Severe	15 (65.2%)	8 (34.7%)	1.875	
Ulcerative Colitis (N = 56)	Remission	7 (70.0%)	3 (30.0%)	2.33	0.3157
	Mild	5 (45.4%)	6 (54.6%)	0.83	
	Moderate	5 (45.4%)	6 (54.6%)	0.83	
	Severe	17 (70.8%)	7 (29.2%)	2.43	

PCR-polymerase chain reaction.

Table 6. Stool inflammatory marker levels in IBD patients with and without enteric infections identified on stool PCR

Inflammatory Marker	Negative Stool PCR	Positive Stool PCR	T-Test *P*-Value	Chi-Square *P*-Value
Lactoferrin (n = 23) - Mean (± stdev)	190 (± 332.52)	468.41 (± 532.66)	0.068	0.017
Calprotectin (n = 5) - Mean (± stdev)	1004.56 (± 720.73)	658.25 (± 587.07)	0.9455	

PCR-polymerase chain reaction.

and positive PCR result was not statistically significant in this subset (P = 0.3157) (**Table 5**).

Inflammatory stool markers and stool PCR positivity

Stool lactoferrin values were also evaluated in a subset analysis: 23 patients had a lactoferrin value. An elevated lactoferrin value of 468.4 (± 532.66) was associated with a positive stool PCR (P = 0.068). A Chi-square *p*-value did show a statistically significant association between lactoferrin and stool PCR positivity; however, the overall number of patients evaluated in this sub-analysis was small relative to the other parts of the study (**Table 6**).

Disease location and stool PCR positivity

Patients were then categorized using the Montreal Classification to analyze disease distribution with respect to stool PCR results (**Table 7**). In patients with ulcerative colitis, those with any distribution of UC, i.e., either E1, E2 or E3, were more likely to be positive than negative for an enteric infection. Patients with pancolitis (E3) were almost three times more likely to have enteric infections than not (OR = 2.8), as 63 (39.4%) had a positive stool PCR, and 23 (8.2%) had a negative PCR. Among patients with left-sided or distal colitis, 11 (3.9%) had a positive PCR, while nine (3.2%) had a negative PCR (OR = 1.2). Among patients with ulcerative proctitis only, five (1.8%) had a positive PCR, and two (0.7%) had a negative PCR (OR = 2.5).

In patients with ileocolonic Crohn's disease (L3), 54 (19.4%) had a positive stool PCR, and 52 (18.6%) had a negative PCR (OR = 1.0); in patients with colonic disease only (L2), 12 (4.3%) patients had a positive stool PCR, and 11 (3.9%) had a negative stool PCR (OR = 1.1). Chi-square *p*-value was also calculated (P = 0.0000897) and indicated differential distribution of PCR positivity within disease subtype. Comparatively, in patients in the CD cohort with ileal disease only (L1), five (1.8%) patients had a positive stool PCR, and 16 (5.7%) had a negative PCR test (OR = 0.31), which indicates that patients were less likely to have a positive stool PCR if they had ileal disease alone. In patients

Table 7. PCR stool positivity by location of disease as based on the Montreal Classification for IBD

Montreal Classification	Positive Stool PCR (N = 167)	Negative Stool PCR (N = 117)	Odds Ratio	Chi-Square *P*-Value
E1	5 (1.8%)	2 (0.7%)	2.5	0.0000897
E2	11 (3.9%)	9 (3.2%)	1.2	
E3	64 (39.4%)	23 (8.2%)	2.8	
L1	5 (1.8%)	16 (5.7%)	0.31	
L2	12 (4.3%)	11 (3.9%)	1.1	
L2+L4	4 (1.4%)	1 (0.4%)	4	
L3	54 (19.4%)	52 (18.6%)	1.0	
L3+L4	10 (3.5%)	0 (0.0%)	Infinite	

PCR-polymerase chain reaction; IBD-inflammatory bowel disease; L1-terminal ileum; L2-colon; L3-ileocolon; L4-upper GI modifier, proximal disease with distal disease; E1-ulcerative proctitis; E2-left-sided UC, distal colitis; E3-pancolitis.

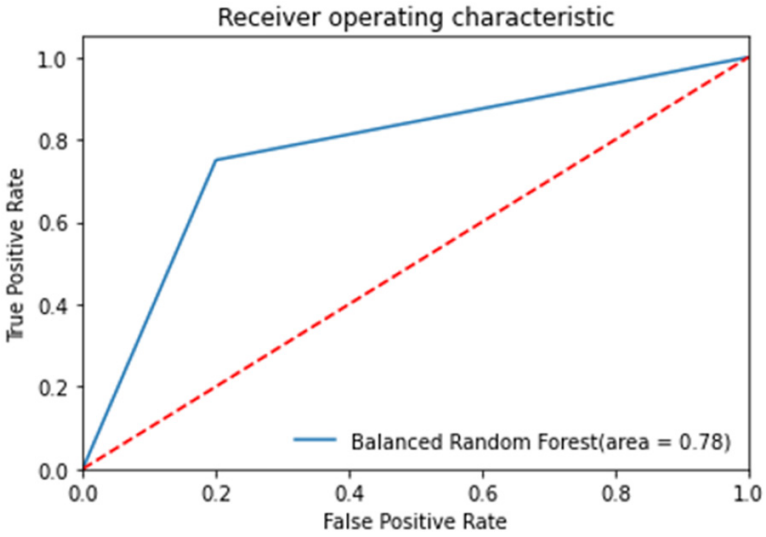

Figure 1. Receiver operating characteristic (ROC) curve for balanced random forest model with target variable of GI panel positivity. Included variables were age, male sex, ulcerative colitis subtype, E3 Montreal classification, L2 Montreal classification and Endoscopy remission.

with colonic and upper GI involvement (L2+L4), four (n = 1.4%) had a positive stool PCR, while only one (0.4%) had a negative PCR test (OR = 4); in patients with ileocolonic disease and upper GI involvement (L3+L4), 10 (3.5%) patients had a positive stool PCR, and no patients had a negative PCR test (OR = infinite).

Multivariate analysis

The number of patients with an endoscopy completed within three months was 151. After removing collinear variables, we included the following variables in model training and evaluation: age, male sex, ulcerative colitis subtype, E3 and L2 Montreal classification, and endoscopy remission. Fecal lactoferrin was not used in the analysis due to only 32 of the 151 patients having this data. After input to a balanced random forest classifier with 10-fold cross-validation, the model reported a mean f1 score of 0.727, mean recall of 0.740, and mean precision of 0.737. An ROC curve demonstrated a mean area under the curve (AUC) of 0.78 (**Figure 1**). Gini importances were 0.75 for age, 0.06 for endoscopic remission, 0.05 for male sex, 0.05 for E3 Montreal classification, 0.05 for L2 Montreal classification, and 0.04 for ulcerative colitis subtype.

Discussion

In this retrospective cohort study, we analyzed the characteristics of IBD patients with enteric infections and associations with disease activity; in addition, we conducted a multivariate analysis using machine learning techniques to identify variables that predict GI pathogen panel positivity, which is still an evolving field of research interest.

The factors that were significantly associated with having a positive stool PCR test included: having a diagnosis of ulcerative colitis, colonic Crohn's involvement, increased endoscopic disease severity and age of 20-50 years. We

found enteric infections in more than half of the patients who had a stool PCR collected. The two most commonly identified enteric infections were *C. difficile* and *E. coli* subtypes.

Based on our analyses, patients with ulcerative colitis were more likely to have a positive stool PCR than those with Crohn's disease. Given the location of inflammation in UC, we suspect that these patients had more breakdown in the colonic mucosal barrier and therefore, higher incidence of enteric infections [10]. Our analysis of disease location using the Montreal Classification showed that any distribution of ulcerative colitis was still more likely to have a positive than negative stool PCR for enteric infections, although patients with pancolitis (E3) were almost three times more likely to have a positive stool PCR, lending credence to our hypothesis that more surface area of potential mucosal breakdown lends a higher risk of enteric infection. Of note, there were fewer patients in the proctitis (E1) and L-sided colitis (E2) than in the pancolitis (E3) cohort, which may have affected this result. In addition, patients with Crohn's disease exhibited a higher likelihood of contracting enteric infections if they had any colonic involvement of their disease, i.e., either ileocolonic or colonic disease. Indeed, the overall data in the Crohn's disease subset of patients also supports the hypothesis that enteric infections correlate with disruption of the mucosal epithelium located in the colonic areas.

Activity of disease as determined by endoscopy within three months of the collected stool PCR was also important; severe disease portended a higher likelihood of enteric infections, regardless of the subtype of IBD, and remission was associated with lower likelihood of enteric infections. A three-month endoscopy window was chosen based on the traditional practice of avoiding colonoscopy at the time of an active enteric infection. Data for a smaller window such as less than four weeks would be limited; most IBD patients with infections identified on stool PCR testing are first treated for the infection prior to undergoing endoscopic evaluation. Sub-analyses of both Crohn's disease and ulcerative colitis confirmed an association between high disease severity, noted on colonoscopy, and the presence of enteric infection. Immune dysregulation manifested by severe inflammatory bowel disease activity may portend a higher risk of infection, and this highlights the importance of achieving mucosal remission.

Additionally, we analyzed stool markers of inflammation in our cohort of IBD patients with enteric infections, specifically lactoferrin. Fecal lactoferrin, a major component of granules released upon neutrophil activation and degranulation, is resistant to degradation and is useful in evaluating inflammation in the GI tract. Fecal lactoferrin has been shown to correlate with levels of disease activity in IBD patients [22-25]. In our study, a positive stool PCR was significantly associated with elevated lactoferrin, which suggests ongoing cell damage and active inflammation in these patients.

Lastly, our multivariate analysis demonstrated the potential use of machine learning techniques to predict GI panel positivity. By considering the complex multidimensional interactions between inputs, our balanced random forest classifier demonstrated an AUC of 0.78. This was even without the inclusion of relevant inflammatory biomarkers such as fecal calprotectin or fecal lactoferrin. Our model demonstrated the significance of age in predicting GI panel positivity in IBD patients, evidenced by a Gini importance of 0.75, markedly higher than the next significant variable, endoscopic remission, with a Gini importance of 0.06.

There were several limitations to this study. First, its retrospective design introduces confounding bias compared to a prospective design. The study was conducted at a single academic IBD referral center, so our findings may not be applicable to smaller hospitals and community settings. The limited sample size in certain sub-analyses may have influenced some of the results. For instance, the small number of patients in the UC sub-analysis of endoscopic disease activity and PCR stool positivity could explain why more patients in remission had a positive PCR. A larger number of patients in future studies will also allow for more robust multivariate analyses utilizing the above machine learning techniques.

In our study, almost 60% of patients had a positive stool PCR. Our GI pathogen positivity rate is higher than previously reported in IBD patients in the literature [4, 13, 15, 16]. This could be

related to our standard of early assessment of patients when presenting with any flare symptoms including acute diarrhea, increased urgency or frequency of bowel movements and having easy accessibility for testing with kits. In fact, send-out kits are readily available at our institution and are sent to patients' homes to improve overall compliance with treatment recommendations. The most common infections identified in our patient population were *C. difficile* and *Enteropathogenic E. coli*, which is consistent with prior studies as well [4, 14, 15]. Our institution has a policy of not accepting formed stools and excluding patients who have taken laxatives in the week prior from C. difficile PCR, antigen testing, or toxin assays. These practices could potentially reduce the identification of patients colonized with C. difficile. As a result of PCR positivity, adjustments were made to change therapy based on endoscopic findings; for example, if significant active disease was found, the patient generally had escalation of medical therapy, and if inactive disease was encountered, less aggressive management related to their IBD therapies was pursued. Further subset analyses of specific changes in medical therapy are of interest.

Our restricted sample size most likely complicated calibration of our predictive model. These limitations were mitigated with the use of tree-based models with a rigorous nested cross-validation framework that minimized the variance of our outcomes. Future analyses with larger sample size and other variables such as inflammatory markers, or IBD clinical scores could further build on the proof of concept provided in this paper.

Conclusion

Overall, our study highlights the importance of achieving endoscopic remission, as this could reduce the likelihood of enteric infections. In addition, we present a novel use of machine learning techniques to conduct a multivariate analysis that identifies factors that predict stool PCR positivity. Further studies are needed to evaluate the importance of histologic remission, the impact of infections on the long-term course of IBD, further roles of stool testing with both molecular pathogen panels and markers, changes made to the IBD treatment and its impact on the incidence of enteric infections in this cohort and the relationship between specific disease distributions and outcomes.

Acknowledgements

The authors thank Drs. Jonathan Feinberg and James Kasper for editing the manuscript.

Written informed consent was obtained from all subjects.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Christopher Kim, Division of Gastroenterology, Department of Medicine, Houston Methodist Hospital, 6550 Fannin St, Smith Tower, Suite 1201, Houston, TX 77030, USA. Tel: 713-441-3372; E-mail: cjintaekim@houstonmethodist.org

References

[1] Alatab S, Sepanlou SG, Ikuta K, Vahedi H, Bisignano C, Safiri S, Sadeghi A, Nixon MR, Abdoli A, Abolhassani H, Alipour V, Almadi MAH, Almasi-Hashiani A, Anushiravani A, Arabloo J, Atique S, Awasthi A, Badawi A, Baig AAA, Bhala N, Bijani A, Biondi A, Borzì AM, Burke KE, Carvalho F, Daryani A, Dubey M, Eftekhari A, Fernandes E, Fernandes JC, Fischer F, Haj-Mirzaian A, Haj-Mirzaian A, Hasanzadeh A, Hashemian M, Hay SI, Hoang CL, Househ M, Ilesanmi OS, Balalami NJ, James SL, Kengne AP, Malekzadeh MM, Merat S, Meretoja TJ, Mestrovic T, Mirrakhimov EM, Mirzaei H, Mohammad KA, Mokdad AH, Monasta L, Negoi I, Nguyen TH, Nguyen CT, Pourshams A, Poustchi H, Rabiee M, Rabiee N, Ramezanzadeh K, Rawaf DL, Rawaf S, Rezaei N, Robinson SR, Ronfani L, Saxena S, Sepehrimanesh M, Shaikh MA, Sharafi Z, Sharif M, Siabani S, Sima AR, Singh JA, Soheili A, Sotoudehmanesh R, Suleria HAR, Tesfay BE, Tran B, Tsoi D, Vacante M, Wondmieneh AB, Zarghi A, Zhang ZJ, Dirac M, Malekzadeh R and Naghavi M. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol 2020; 5: 17-30.

[2] Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, Panaccione R, Ghosh S, Wu JCY, Chan FKL, Sung JJY and Kaplan GG. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 2017; 390: 2769-2778.

[3] Lewis JD, Parlett LE, Jonsson Funk ML, Brensinger C, Pate V, Wu Q, Dawwas GK, Weiss A, Constant BD, McCauley M, Haynes K, Yang JY, Schaubel DE, Hurtado-Lorenzo A and Kappel-

man MD. Incidence, prevalence, and racial and ethnic distribution of inflammatory bowel disease in the United States. Gastroenterology 2023; 165: 1197-1205, e2.
[4] Axelrad JE, Joelson A, Green PHR, Lawlor G, Lichtiger S, Cadwell K and Lebwohl B. Enteric infections are common in patients with flares of inflammatory bowel disease. Am J Gastroenterol 2018; 113: 1530-1539.
[5] Masclee GM, Penders J, Pierik M, Wolffs P and Jonkers D. Enteropathogenic viruses: triggers for exacerbation in IBD? A prospective cohort study using real-time quantitative polymerase chain reaction. Inflamm Bowel Dis 2013; 19: 124-31.
[6] Micic D, Hirsch A, Setia N and Rubin DT. Enteric infections complicating ulcerative colitis. Intest Res 2018; 16: 489-493.
[7] Axelrad JE, Joelson A, Nobel YR, Lawlor G, Green PHR, Lichtiger S and Lebwohl B. Enteric infection in relapse of inflammatory bowel disease: the utility of stool microbial PCR testing. Inflamm Bowel Dis 2017; 23: 1034-1039.
[8] Porter CK, Tribble DR, Aliaga PA, Halvorson HA and Riddle MS. Infectious gastroenteritis and risk of developing inflammatory bowel disease. Gastroenterology 2008; 135: 781-786.
[9] Garcia Rodriguez LA, Ruigomez A and Panes J. Acute gastroenteritis is followed by an increased risk of inflammatory bowel disease. Gastroenterology 2006; 130: 1588-94.
[10] Shea-Donohue T, Fasano A, Smith A and Zhao A. Enteric pathogens and gut function: role of cytokines and STATs. Gut Microbes 2010; 1: 316-324.
[11] Tarris G, de Rougemont A, Charkaoui M, Michiels C, Martin L and Belliot G. Enteric viruses and inflammatory bowel disease. Viruses 2021; 13: 104.
[12] Nitzan O, Elias M, Chazan B, Raz R and Saliba W. Clostridium difficile and inflammatory bowel disease: role in pathogenesis and implications in treatment. World J Gastroenterol 2013; 19: 7577-85.
[13] Binion D. Clostridium difficile infection and inflammatory bowel disease. Gastroenterol Hepatol (N Y) 2016; 12: 334-7.
[14] Ahmad W, Nguyen NH, Boland BS, Dulai PS, Pride DT, Bouland D, Sandborn WJ and Singh S. Comparison of multiplex gastrointestinal pathogen panel and conventional stool testing for evaluation of diarrhea in patients with inflammatory bowel diseases. Dig Dis Sci 2019; 64: 382-390.
[15] Ahn JS, Seo SI, Kim J, Kim T, Kang JG, Kim HS, Shin WG, Jang MK and Kim HY. Efficacy of stool multiplex polymerase chain reaction assay in adult patients with acute infectious diarrhea. World J Clin Cases 2020; 8: 3708-3717.
[16] Issa M, Ananthakrishnan AN and Binion DG. Clostridium difficile and inflammatory bowel disease. Inflamm Bowel Dis 2008; 14: 1432-42.
[17] Mylonaki M, Langmead L, Pantes A, Johnson F and Rampton DS. Enteric infection in relapse of inflammatory bowel disease: importance of microbiological examination of stool. Eur J Gastroenterol Hepatol 2004; 16: 775-8.
[18] Seyed Tabib NS, Madgwick M, Sudhakar P, Verstockt B, Korcsmaros T and Vermeire S. Big data in IBD: big progress for clinical practice. Gut 2020; 69: 1520-1532.
[19] Panchavati S, Zelin NS, Garikipati A, Pellegrini E, Iqbal Z, Barnes G, Hoffman J, Calvert J, Mao Q and Das R. A comparative analysis of machine learning approaches to predict C. difficile infection in hospitalized patients. Am J Infect Control 2022; 50: 250-257.
[20] Yuan F, Zhang YH, Kong XY and Cai YD. Identification of candidate genes related to inflammatory bowel disease using minimum redundancy maximum relevance, incremental feature selection, and the shortest-path approach. Biomed Res Int 2017; 2017: 5741948.
[21] Toubeau J, Bottieau J, Vallée F and De Grève Z. Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets. IEEE Trans Power Syst 2019; 34: 1203-1215.
[22] Kane SV, Sandborn WJ, Rufo PA, Zholudev A, Boone J, Lyerly D, Camilleri M and Hanauer SB. Fecal lactoferrin is a sensitive and specific marker in identifying intestinal inflammation. Am J Gastroenterol 2003; 98: 1309-14.
[23] Sugi K, Saitoh O, Hirata I and Katsu K. Fecal lactoferrin as a marker for disease activity in inflammatory bowel disease: comparison with other neutrophil-derived proteins. Am J Gastroenterol 1996; 91: 927-34.
[24] Walker TR, Land ML, Cook TM, Boone JH, Lyerly D and Rufo PA. Serial fecal lactoferrin measurements are useful in the interval assessment of patients with active and inactive inflammatory bowel disease. Gastroenterology 2004; 126: A215.
[25] Abraham BP and Kane S. Fecal markers: calprotectin and lactoferrin. Gastroenterol Clin North Am 2012; 41: 483-95.