Title

Machine Learning-based Characterization of Longitudinal Health Care Utilization Among Patients With Inflammatory Bowel Diseases.

Journal

Inflammatory Bowel Diseases, 30(5)

Authors

Limketkai, Berkeley
Maas, Laura
Krishna, Mahesh
et al.

Publication Date

2024-05-02

DOI

10.1093/ibd/izad127

Peer reviewed

Inflammatory Bowel Diseases, 2024, **30**, 697–703
https://doi.org/10.1093/ibd/izad127
Advance access publication 15 July 2023
Leading Off

Machine Learning-based Characterization of Longitudinal Health Care Utilization Among Patients With Inflammatory Bowel Diseases

Berkeley N. Limketkai, MD, PhD,[*] Laura Maas, MD,[†] Mahesh Krishna, BA,[†] Anoushka Dua, MD,[*] Lauren DeDecker, MD,[*] Jenny S. Sauk, MD,[*] and Alyssa M. Parian, MD[†]

From the [*]Center for Inflammatory Bowel Diseases, Vatche and Tamar Manoukian Division of Digestive Diseases, UCLA School of Medicine, Los Angeles, CA, USA
[†]Division of Gastroenterology and Hepatology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Address correspondence to: Berkeley N. Limketkai, MD, PhD, 100 UCLA Medical Center Plaza, Suite 345, Los Angeles, California 90095, USA (berkeley.limketkai@gmail.com).

Background: Inflammatory bowel disease (IBD) is associated with increased health care utilization. Forecasting of high resource utilizers could improve resource allocation. In this study, we aimed to develop machine learning models (1) to cluster patients according to clinical utilization patterns and (2) to predict longitudinal utilization patterns based on readily available baseline clinical characteristics.

Methods: We conducted a retrospective study of adults with IBD at 2 academic centers between 2015 and 2021. Outcomes included different clinical encounters, new prescriptions of corticosteroids, and initiation of biologic therapy. Machine learning models were developed to characterize health care utilization. Poisson regression compared frequencies of clinical encounters.

Results: A total of 1174 IBD patients were followed for more than 5673 12-month observational windows. The clustering method separated patients according to low, medium, and high resource utilizers. In Poisson regression models, compared with low resource utilizers, moderate and high resource utilizers had significantly higher rates of each encounter type. Comparing moderate and high resource utilizers, the latter had greater utilization of each encounter type, except for telephone encounters and biologic therapy initiation. Machine learning models predicted longitudinal health care utilization with 81% to 85% accuracy (area under the receiver operating characteristic curve 0.84-0.90); these were superior to ordinal regression and random choice methods.

Conclusion: Machine learning models were able to cluster individuals according to relative health care resource utilization and to accurately predict longitudinal resource utilization using baseline clinical factors. Integration of such models into the electronic medical records could provide a powerful semiautomated tool to guide patient risk assessment, targeted care coordination, and more efficient resource allocation.

Key Words: machine learning, health care utilization, quality improvement

Introduction

Inflammatory bowel disease (IBD) characteristically presents with a relapsing and remitting course, while often requiring the use of powerful immunosuppressants that do not necessarily provide complete or durable control of intestinal inflammation.[1,2] Patients with IBD may subsequently experience disease progression and disease-related complications that pose significant morbidity and increased health care utilization.[3–5] These patients generally experience increased need for communication with their gastroenterology providers, laboratory and other diagnostic testing, emergency department (ED) visits, hospitalizations, and surgery. Health care utilization patterns may nonetheless fluctuate within individuals and differ widely across individuals with IBD. An individual with severely active disease might require intense monitoring, medication adjustments, hospitalization, and even surgery. The same individual, when in remission, might not require follow-up for many months. Although some individuals experience an overall mild disease course, others may experience severe disease requiring hospitalization and surgery.

Increased health care utilization can often strain the operational capacity and resources of ambulatory practices. Given the highly variable nature of health care utilization within and across individuals over time, semiautomated methods that facilitate identification and forecasting of potential high-resource utilizers could promote more efficient and dynamic resource allocation for proactive disease monitoring, timely treatment, and clinical staffing in ambulatory practices and health care systems. Focused interventions could also help reduce disease-related complications and overall costs of care. As such, this study aimed to develop (1) a machine learning model to cluster patients according to health care utilization patterns; and (2) another model to predict longitudinal utilization patterns based on readily available clinical characteristics at baseline.

Methods

Data Source

In this retrospective cohort study, the source population included adults (ages 18 years or older) with IBD

Received for publication: January 23, 2023. Editorial Decision: June 4, 2023

Key Messages

What is already known?

Inflammatory bowel disease is associated with increased health care utilization, and improved forecasting of high resource utilizers could facilitate more efficient resource allocation.

What is new here?

Machine learning methods can be used to characterize longitudinal health care utilization using readily available baseline clinical data.

How can this study help patient care?

Integration of these models into the electronic medical records could provide a powerful semiautomated tool to guide patient risk assessment, targeted care coordination, and more efficient resource allocation.

who had medical visits in gastroenterology clinics at 2 large academic centers between January 2015 and May 2021. Potential patients were initially identified using the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes 555 or 556 and the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes K50 or K51 in any diagnosis position at the baseline gastroenterology visit. Diagnoses were confirmed based on available endoscopic data and/or consistency of IBD diagnosis in subsequent clinical encounters. Patients who did not have at least 2 IBD-related encounters, who did not have at least 12 months of follow-up, and whose records did not contain minimum IBD-related descriptors in the clinic notes were excluded.

Data on patient demographics (age, sex, race, ethnicity), anthropometrics (height, weight), IBD type, disease characteristics (behavior, location, severity), gastrointestinal symptoms, smoking status (current, former, never), medications, and surgical history were abstracted. Disease activity measures were recorded as the Harvey-Bradshaw Index (HBI) for Crohn's disease (CD) and as the Simple Clinical Colitis Activity Index (SCCAI) for ulcerative colitis (UC). Data on extraintestinal manifestations were not included due to inconsistent recording in the clinic notes among gastroenterology providers who did not subspecialize in IBD. Laboratory values considered for inclusion are listed in Supplementary Table 1. Laboratory tests with >30% missing data were excluded to reduce potential introduction of bias and due to lower likelihood of test data availability in future clinical application of the machine learning models. Despite not meeting these criteria, calprotectin was included due to its strong correlation with disease activity, and erythrocyte sedimentation rate was excluded due to significant collinearity with the already included C-reactive protein (CRP). The final list of laboratory tests included complete blood count, comprehensive metabolic panel, CRP, and calprotectin. Ten-fold multiple imputation was used to address missing data in these tests.

The study was approved by the institutional review boards of the University of California, Los Angeles (UCLA) and Johns Hopkins University.

Outcomes

Clinical outcomes included the frequency of clinic visits, telephone or electronic communication, ED visits, hospitalizations, radiology tests, endoscopic procedures, IBD-related surgeries, initiation of biologic therapy, and initiation of corticosteroids within 12 months of the baseline visit. Clinic visits included all in-person and telemedicine ambulatory encounters in the gastroenterology and colorectal surgery clinics. Communication-related encounters included all telephone calls or electronic messages recorded by the gastroenterology or colorectal surgery divisions in the electronic medical records (EMRs). Due to limitations of how these data were structured, each conversation thread could appear as only 1 encounter even if it contained multiple telephone calls or electronic messages. ED visits and hospitalizations were counted for all indications. Endoscopic procedures included upper endoscopies, small bowel enteroscopies, colonoscopies, and pouchoscopies. Surgeries were limited to bowel resection surgeries. Radiologic tests included radiographs, computed tomography, magnetic resonance imaging, ultrasound, and nuclear medicine studies.

Resource Utilization Clustering

The objective of the first machine learning model was to meaningfully distill the aforementioned clinical outcomes into a composite ordinal measure of health care utilization using previously applied unsupervised methods.[6] The earliest baseline visit for each patient was the first recorded gastroenterology clinic visit with a diagnosis of IBD. Starting at the earliest baseline visit for each patient, rolling observational data windows were then generated based on available gastroenterology clinic visits with 12 months of clinical follow-up. Data preprocessing with z-score normalization was applied to each 12-month encounter type. Unsupervised 10-fold k-means clustering was then used to separate patients into low, moderate, and high resource utilization patterns. The delineation of utilization patterns as tertiles was based on the intention to provide clinically meaningful and actionable risk stratification, as similarly performed in other studies.[7–9] Each initial centroid was randomly generated and repeated across 10 iterations of the algorithm. To subsequently evaluate the validity of assignments, Poisson regression was used to compare the frequency of each type of clinical encounter without data preprocessing across resource utilization clusters.

Resource Utilization Prediction

For development of the machine learning model to predict longitudinal health care utilization patterns, the training set was generated from a randomly sampled 80% of observational windows; the validation set comprised the remaining 20%. Exposures were standardized baseline clinical data, and the outcomes were the previously assigned clusters of low, moderate, or high resource utilization. Machine learning models were trained using random forest classification (RFC) or extreme gradient boosting (XGB), tuned by grid-searching parameters applicable to each algorithm with stratified 10-fold cross-validation. The models were subsequently validated with the validation set. Two sets of models included features with and without laboratory test results. Precision, recall, F1 score, accuracy, and area under the receiver operating characteristic curve (AUROC) were estimated for each model. Model performances were also compared with ordinal regression and weighted random classification (ie, random number generation with possible options weighted according to relative prevalence of each category).

Statistical Analysis

Categorical variables were compared using the χ^2 test, and continuous variables were compared using the Student t test. Statistical significance was defined as a 2-tailed α threshold of <0.05. All statistical analyses and machine learning model development were performed using Python 3.10.

Results

Patient Characteristics

There were 1174 IBD (42.4% CD, 57.6% UC) patients whose follow-up data comprised at least 5673 12-month observational windows. The mean age at baseline across observational windows was 42.1 years (standard deviation [SD] 16.3), and most individuals were female (53.6%), white (74.5%), non-Hispanic (91.4%), and nonsmoking (94.9%). Patients with CD more commonly presented with an inflammatory behavior (43.8%) and ileocolic location (46.6%, Table 1). For UC, most patients (53.6%) had inflammation beyond the splenic flexure. Compared with patients with UC, fewer patients with CD were on mesalamine (15.4% vs 57.9%, $P < .01$), and more were on immunomodulators (36.7% vs 33.5%, $P = .02$) and biologic therapy (46.0% vs 33.7%, $P < .01$). More patients with CD were in clinical remission (54.5% vs 49.2%, $P < .01$).

Resource Utilization Clustering

Unsupervised methods identified 3 discrete resource utilization clusters, labeled as low, moderate, and high resource utilization. Low resource utilizers had the lowest mean 12-month frequency of each measured encounter type (Table 2, Figure 1). In Poisson regression models, compared with low resource utilizers, moderate and heavy resource utilizers had significantly higher rates of each encounter type. Comparing moderate with high resource utilizers, the latter had greater utilization of each encounter type, except for telephone encounters (−0.04; 95% confidence interval [CI], −0.08 to −0.00; $P = .03$) and biologic therapy initiation (−0.43; 95% CI, −0.63 to −0.24; $P < .01$).

Resource Utilization Prediction

The machine learning algorithms performed well when using baseline clinical data to predict 12-month resource utilization category (Table 3, Figure 2). Validation AUROC for RFC and XGB models with laboratory features were 0.90 and 0.89, respectively. Without laboratory features, AUROC were slightly lower at 0.87 and 0.84, respectively. Accuracy was identical at 0.85 for both algorithms with laboratory features and similar (RFC 0.82; XGB 0.81) for both algorithms without laboratory features. By contrast, ordinal regression methods and weighted random classification had inferior performance. AUROC for ordinal regression was 0.76 with laboratory features and 0.72 without laboratory features. Accuracy for ordinal regression was 0.75 with and without laboratory features, whereas accuracy for weighted random classification was 0.60 (unweighted or simple random classification was 0.34).

Table 1. Clinical baseline characteristics across observational windows.

	CD (n = 2809)	UC (n = 2203)	P
Age, mean years (SD)	40.9 (16.0)	43.6 (16.5)	< 0.01
Sex (%)			0.83
Male	1308 (46.6)	1020 (46.3)	
Female	1501 (53.4)	1183 (53.7)	
Race			< 0.01
White	2183 (77.7)	1552 (70.4)	
Black	170 (6.0)	92 (4.2)	
Asian or Pacific Islander	93 (3.3)	190 (8.6)	
Multiple races	35 (1.2)	57 (2.6)	
Other	328 (11.7)	312 (14.2)	
Hispanic ethnicity	193 (6.9)	237 (10.8)	< 0.01
Smoking status			< 0.01
Never	2016 (71.8)	1608 (73.0)	
Former	601 (21.4)	531 (24.1)	
Current	192 (6.8)	64 (2.9)	
CD behavior			
Inflammatory	1229 (43.8)	N/A	
Stricturing	793 (28.2)	N/A	
Penetrating	338 (12.0)	N/A	
Unknown	449 (16.0)	N/A	
Perianal disease	123 (4.4)	N/A	
CD location			
Ileal	990 (35.2)	N/A	
Colonic	403 (14.3)	N/A	
Ileocolonic	1310 (46.6)	N/A	
Upper GI involvement	106 (3.8)	N/A	
UC extent			
Rectum	N/A	158 (7.2)	
Left side	N/A	863 (39.2)	
Beyond splenic flexure	N/A	1182 (53.6)	
On mesalamine	432 (15.4)	1275 (57.9)	< 0.01
On immunomodulator	1032 (36.7)	739 (33.5)	0.02
On biologic therapy	1292 (46.0)	742 (33.7)	< 0.01
Disease activity			< 0.01
Remission	15327 (54.5)	1083 (49.2)	
Mild	854 (30.4)	685 (31.1)	
Moderate	330 (11.7)	313 (14.2)	
Severe	93 (3.3)	122 (5.5)	

Abbreviations: CD, Crohn's disease; GI, gastrointestinal; N/A, not applicable; SD, standard deviation; UC, ulcerative colitis.

The most important features for the RFC model with laboratory values were age, albumin, hemoglobin, white blood cell count, platelet count, calcium, and disease activity. The most important features for the RFC model without laboratory values were age, disease activity, current biologic use, prior biologic use, disease location, and race. The most important features for the XGB model with laboratory values were disease activity, current biologic use, prior biologic use, albumin, and calcium. The most important features for the XGB model without laboratory values were disease activity, current biologic use, prior biologic use, ethnicity, and sex.

Table 2. Relative frequency of encounters comparing moderate and heavy resource utilizers with low resource utilizers.

	Low Resource Utilization, Mean Encounters per 12 Months	Moderate Resource Utilization β (95% CI)*	Heavy Resource Utilization β (95% CI)*
Clinic visits	1.52	1.24 (1.20-1.28)	1.64 (1.58-1.69)
Telephone encounters	4.86	1.18 (1.16-1.20)	1.14 (1.10-1.17)
ED encounters	0.09	0.69 (0.51-0.86)	3.31 (3.17-3.44)
Hospital admissions	0.05	1.24 (1.05-1.44)	3.93 (3.77-4.09)
Radiology tests	1.52	0.57 (0.53-0.62)	1.52 (1.46-1.57)
Endoscopic procedures	0.56	0.65 (0.58-0.72)	0.75 (0.62-0.88)
IBD-related surgeries	0.66	0.90 (0.84-0.96)	1.66 (1.57-1.74)
Biologic therapy	0.03	3.30 (3.11-3.50)	2.87 (2.61-3.14)
Corticosteroid therapy	0.60	1.73 (1.68-1.79)	2.46 (2.39-2.52)

Abbreviations: CI, confidence interval; ED, Emergency Department; IBD, inflammatory bowel disease.
*All $P < .001$.

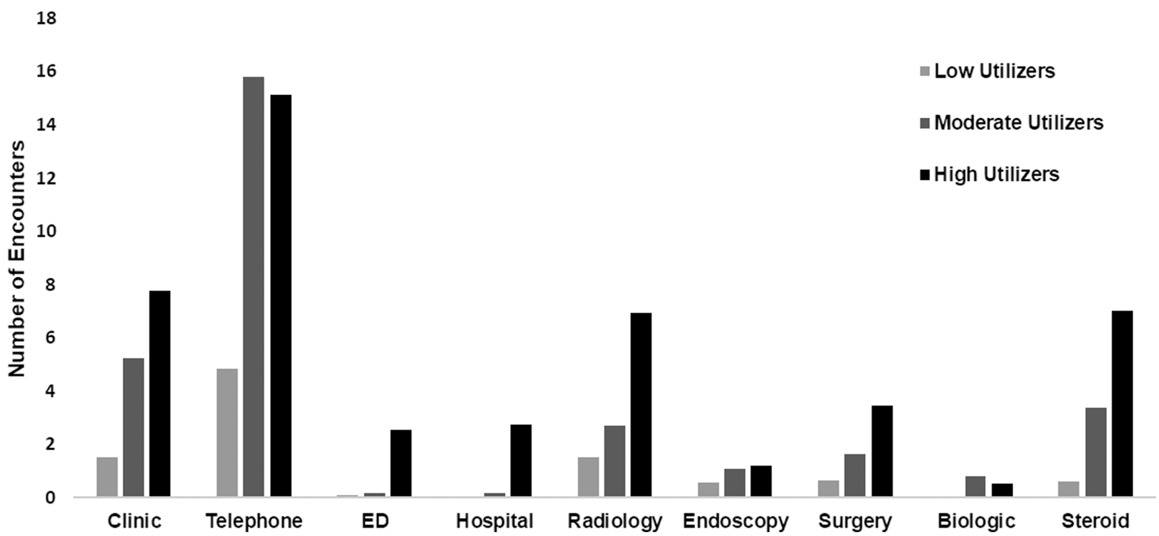

Figure 1. Health care utilization patterns.

Discussion

In this study, we developed a machine learning–based model that successfully stratified individuals according to overall health care utilization. An unsupervised method allowed us to aggregate disparate yet related encounter types for determination of 3-tiered resource utilization patterns. Relative resource utilization demonstrated a stepwise increase in mean frequencies of discrete and overall encounter types across utilization tertiles, reflecting the appropriateness of categorization. We additionally developed machine learning models that examined readily available baseline clinical data to predict relative health care resource utilization with good accuracy.

To predict health care utilization, tree-based algorithms with finely tuned parameters generated similar performances between RFC and XGB models. Their performance was significantly superior to regression-based models and weighted random classification, indicating that they provide added value over traditional prediction models and even chance. A key advantage of machine learning–based methods in these applications is the ability to analyze subtle and multidimensional relationships among a vast array of clinical variables when predicting an outcome. Compared with other machine learning models for outcome prediction, our models of health care utilization performed similarly or modestly better. A prior RFC model developed to predict flares with IBD-related hospitalization and corticosteroid prescriptions as surrogate markers used 20 368 patients in the Veterans Health Administration and had an AUROC of 0.85 in the test cohort.[10] When prior hospitalization or corticosteroid use data were included as features, the AUROC increased to 0.87. An analogous model using 95 878 patients in the Optum Electronic Health Records Database had an AUROC of 0.80 in the test cohort.[11] A potential reason for the decrement in performance when using large administrative data sets is the lack of granular clinical data otherwise available in local cohorts. Unlike these prior RFC models, ours were intended to predict a multiclass outcome of health care utilization and relied on a markedly smaller training set. The similarity in performance nonetheless suggests that nationwide-scaled cohorts may not be mandatory for developing performant models, although larger-scale models could have the advantage of improving generalizability.

Table 3. Model training and validation performance.

	AUROC	Precision	Recall	F1 Score	Accuracy
Weighted random classification	N/A	0.59	0.60	0.59	0.60
Ordinal regression					
With lab values	0.76	0.67	0.75	0.69	0.75
Without lab values	0.72	0.68	0.75	0.69	0.75
Training Performance					
Random forest classification					
With lab values	1.00	0.98	0.98	0.98	0.98
Without lab values	0.96	0.90	0.89	0.88	0.89
Extreme gradient boosting					
With lab values	0.99	0.96	0.96	0.96	0.96
Without lab values	0.96	0.90	0.90	0.89	0.90
Validation Performance					
Random forest classification					
With lab values	0.90	0.85	0.85	0.85	0.85
Without lab values	0.87	0.82	0.82	0.80	0.82
Extreme gradient boosting					
With lab values	0.89	0.85	0.85	0.84	0.85
Without lab values	0.84	0.80	0.81	0.78	0.81

Abbreviations: AUROC, area under the receiver operating characteristic curve; N/A, not applicable.

We tested models with and without laboratory features to assess the feasibility of streamlining the baseline clinical variables needed. There was only a modest decrement in performance when excluding laboratory features, suggesting that a simpler set of baseline clinical characteristics may suffice for accurately predicting longitudinal resource utilization. Disease activity and biologic therapy use were unsurprisingly important features in RFC and XGB models. These 2 factors are expected to correlate with increased need for health care utilization. For laboratory tests, among the most important features were albumin, hemoglobin, and white blood cell count. Albumin is a negative acute phase reactant, which has been associated with worse disease activity.[12,13] Anemia is common among patients with IBD, particularly those with more severe disease.[14,15] Surprisingly, CRP and calprotectin were not leading features of the prediction models. One hypothesis for their attenuated predictive value is collinearity with acute disease activity, which itself emerged among the most important features.[16] By contrast, albumin and anemia are more reflective of longitudinal disease severity and complications.[12–15] Race and ethnicity emerged among the important features for the RFC and XGB models without laboratory data, respectively. This finding suggests a role that racial and ethnic disparities may play in determining clinical outcomes and would be worthwhile for future investigation.[17]

As clinical intuition often suffices to forecast a patient's disease course and eventual resource utilization, the value of machine learning models is to provide discriminant performance in a more efficient and automated platform that can perform at scale. In practical terms, the machine learning algorithm could be integrated into an EMR system without need for extraction of additional data. Some potential benefits could include improved staff resource allocation based on the ratio of high vs low health care utilizers, deployment of a care coordinator who can monitor and assist high-risk patients to reduce need for ED visits and hospitalization, and targeted increase in frequency of follow-up contact with high-risk patients to improve clinical outcomes. Such an application could also be extrapolated to other diseases to improve the quality of care beyond IBD.

There were several strengths of the study. First, the study included a large sample of patients aggregated across multiple gastroenterology clinics and care providers on the East and West coasts of the United States over a span of at least 7 years. This diversity of patients and practices helps improve the generalizability of the models. Second, we tuned several models with and without laboratory features to seek optimal performers. Both RFC and XGB models performed well and consistently in all calculated metrics. Important features that drove the models also had clinical plausibility. Third, the models were trained and validated with separate data sets to avoid artificial inflation of validation accuracy due to excessive goodness of fit. For instance, the AUROC of the XGB model was 0.993 with the training set and 0.887 with the validation set.

There were several limitations of the study. First, while all outcome measures were collected prospectively through standard operating procedures, it was possible to have missing events from encounters outside the health system. However, this was uncommon among patients who had longitudinal follow-up in our clinics for 12 months or more. These events could alter accuracy of estimates for ED visits and hospitalizations, but not significantly, as they comprised the minority of encounter types. Second, our classification model used frequency of different encounter types to determine relative health care utilization. The model ignored duration, severity, or cost of hospitalization. Third, the baseline data may have missed important unmeasured confounders that could have affected the outcome, such as stress and

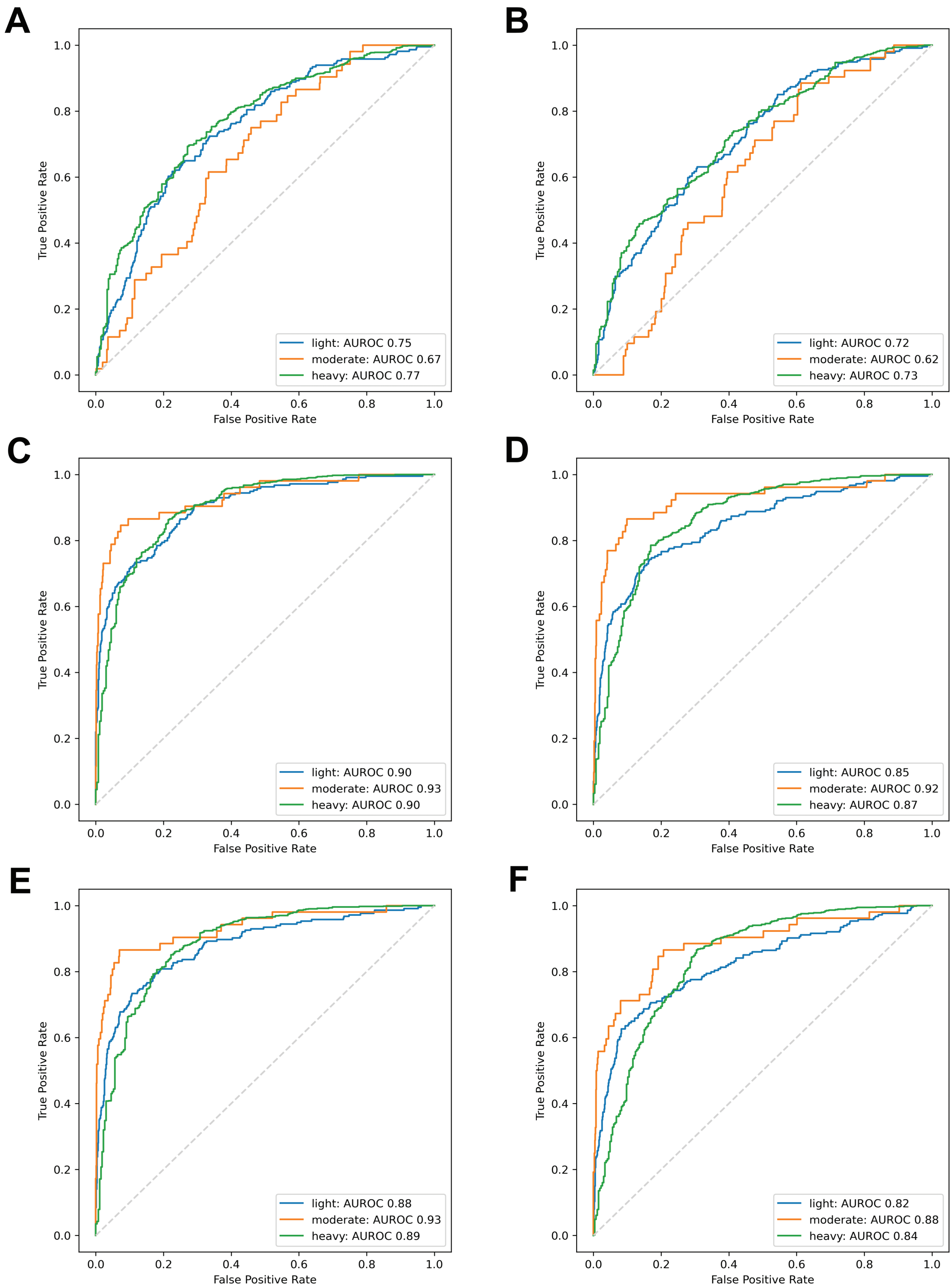

Figure 2. Receiver operating characteristic curves using validation data for (A) ordinal regression with laboratory features, (B) ordinal regression without laboratory features, (C) random forest classification with laboratory features, (D) random forest classification without laboratory features, (E) extreme gradient boosting with laboratory features, and (F) extreme gradient boosting without laboratory features.

anxiety, diet, financial security, insurance coverage, and other social determinants of health. Finally, multiple telephone calls or electronic messages within each conversation thread was only recorded as 1 encounter in the electronic health records. This limitation thus underestimated the true number of calls and messages. However, the number of discrete conversations could still serve as an adequate surrogate measure of resource utilization.

In conclusion, we developed machine learning models to efficiently cluster individuals according to relative health care resource utilization and to accurately predict longitudinal resource utilization based on easily obtained clinical factors at baseline. Model performance was significantly superior to traditional regression methods. Integration of such models into the EMR could provide a powerful semiautomated tool to guide patient risk assessment, targeted care coordination, and more efficient resource allocation.

Supplementary Data

Supplementary data is available at *Inflammatory Bowel Diseases* online.

Acknowledgments

Authors greatly appreciate Dr. Helen Xu's invaluable assistance with data acquisition.

Author Contributions

B.N.L. contributed to the study concept, study design, data acquisition, data interpretation, manuscript preparation, final approval of the manuscript, and study supervision. L.M., M.K., A.D., and L.D. contributed to the study concept, data acquisition, data interpretation, critical review, and final approval of the manuscript. J.S.S. contributed to the data interpretation, critical review, and final approval of the manuscript. A.M.P. contributed to the study concept, study design, data acquisition, data interpretation, critical review, final approval of the manuscript, and study supervision.

Funding

None.

Conflicts of Interest

None.

References

1. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *Lancet*. 2017;389(10080):1741-1755.
2. Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel JF. Ulcerative colitis. *Lancet*. 2017;389(10080):1756-1770.
3. Terlizzi EP, Dahlhamer JM, Xu F, Wheaton AG, Greenlund KJ; Health Care Utilization Among U.S. Adults with inflammatory bowel disease, 2015-2016. *Natl Health Stat Rep*. 2021;(152):1-7.
4. Tsai L, Nguyen NH, Ma C, et al. Systematic review and meta-analysis: risk of hospitalization in patients with ulcerative colitis and Crohn's disease in population-based Cohort studies. *Dig Dis Sci*. 2022;67(6):2451-2461.
5. Solberg IC, Vatn MH, Hoie O, et al.; IBSEN Study Group. Clinical course in Crohn's disease: results of a Norwegian population-based ten-year follow-up study. *Clin Gastroenterol Hepatol*. 2007;5(12):1430-1438.
6. Limketkai BN, Hamideh M, Shah R, Sauk JS, Jaffe N. Dietary patterns and their association with symptoms activity in inflammatory bowel diseases. *Inflamm Bowel Dis*. 2022;28(11):1627-1636.
7. Mumtaz K, Issak A, Porter K, et al. Validation of risk score in predicting early readmissions in decompensated cirrhotic patients: a model based on the administrative database. *Hepatology*. 2019;70(2):630-639.
8. Hill SS, Harnsberger CR, Crawford AS, et al. Creation and institutional validation of a readmission risk calculator for elective colorectal surgery. *Dis Colon Rectum*. 2020;63(10):1436-1445.
9. Crabb BT, Hamrick F, Campbell JM, et al. Machine learning-based analysis and prediction of unplanned 30-day readmissions after pituitary adenoma resection: a multi-institutional retrospective study with external validation. *Neurosurgery*. 2022;91(2):263-271.
10. Waljee AK, Lipson R, Wiitala WL, et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis*. 2017;24(1):45-53.
11. Gan RW, Sun D, Tatro AR, et al. Replicating prediction algorithms for hospitalization and corticosteroid use in patients with inflammatory bowel disease. *PLoS One*. 2021;16(9):e0257520.
12. Chakravarty BJ. Predictors and the rate of medical treatment failure in ulcerative colitis. *Am J Gastroenterol*. 1993;88(6):852-855.
13. Khan N, Patel D, Shah Y, Trivedi C, Yang YX. Albumin as a prognostic marker for ulcerative colitis. *World J Gastroenterol*. 2017;23(45):8008-8016.
14. Koutroubakis IE, Ramos-Rivers C, Regueiro M, et al. Persistent or recurrent anemia is associated with severe and disabling inflammatory bowel disease. *Clin Gastroenterol Hepatol*. 2015;13(10):1760-1766.
15. Maas LA, Krishna M, Parian AM. Ironing it all out: a comprehensive review of iron deficiency anemia in inflammatory bowel disease patients. *Dig Dis Sci*. 2023;68(2):357-369.
16. Sakurai T, Saruta M. Positioning and usefulness of biomarkers in inflammatory bowel disease. *Digestion*. 2023;104(1):30-41.
17. Barnes EL, Loftus EV, Jr, Kappelman MD. Effects of race and ethnicity on diagnosis and management of inflammatory bowel diseases. *Gastroenterology*. 2021;160(3):677-689.