Machine Learning Based Prediction of Incident Cases of Crohn's Disease Using Electronic Health Records from a Large Integrated Health System

Julian Hugo[1*], Susanne Ibing[1,2*(✉)], Florian Borchert[1], Jan Philipp Sachs[1,2], Judy Cho[3], Ryan C. Ungaro[4], and Erwin P. Böttinger[1,2]

[1] Digital Health Center, Hasso Plattner Institute,
University of Potsdam, Potsdam, Germany
`Susanne.Ibing@hpi.de`

[2] Hasso Plattner Institute for Digital Health at Mount Sinai,
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3] The Charles Bronfman Institute for Personalized Medicine,
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4] The Henry D. Janowitz Division of Gastroenterology,
Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract. Early diagnosis and treatment of Crohn's Disease (CD) is associated with decreased risk of surgery and complications. However, diagnostic delay is common in clinical practice. In order to better understand CD risk factors and disease indicators, we identified incident CD patients and controls within the Mount Sinai Data Warehouse (MSDW) and developed machine learning (ML) models for disease prediction.
CD incident cases were defined based on CD diagnosis codes, medication prescriptions, healthcare utilization before first CD diagnosis, and clinical text, using structured Electronic Health Records (EHR) and clinical notes from MSDW. Cases were matched to controls based on sex, age and healthcare utilization. Thus, we identified 249 incident CD cases and 1,242 matched controls in MSDW. We excluded data from 180 days before first CD diagnosis for cohort characterization and predictive modeling. Clinical text was encoded by term frequency-inverse document frequency and structured EHR features were aggregated. We compared three ML models: Logistic Regression, Random Forest, and XGBoost.
Gastrointestinal symptoms, for instance anal fistula and irritable bowel syndrome, are significantly overrepresented in cases at least 180 days before the first CD code (prevalence of 33% in cases compared to 12% in controls). XGBoost is the best performing model to predict CD with an AUROC of 0.72 based on structured EHR data only. Features with highest predictive importance from structured EHR include anemia lab values and race (white). The results suggest that ML algorithms could enable earlier diagnosis of CD and reduce the diagnostic delay.

Keywords: Crohn disease · Diagnostic delay · Electronic health records

[*] These authors contributed equally to this work.

1 Introduction

Inflammatory bowel disease (IBD) with its main entities Crohn's disease (CD) and ulcerative colitis (UC) comprises a group of chronic immune-mediated diseases of the gastrointestinal (GI) tract with relapsing disease course [18,19].

Diagnostic delay, the time between initial manifestation of symptoms and clinical diagnosis of a disease, is a common problem in IBD.

According to a recent meta-analysis by Jayasooriya et al., the median diagnostic delay in CD is 8.0 months (6.2 months in high-income countries), compared to a significantly shorter time period of 3.7 months in UC. Diagnostic delay in CD increases the risk of complications, such as major surgery, strictures, and penetrating disease [12]. Danese et al. [8] described the development and validation of a 'Red Flags Index', a diagnostic tool comprised of 21 symptoms and signs suggestive of CD that, according to the authors, cannot be applied to general CD screening, however, possibly can serve as support tool to prioritize patients for fecal calprotectin (FC) screening [9]. Across individual studies, the identified risk factors varied and did not result in consistent patient features predictive of prolonged time to diagnosis [12]. There is a need to reduce diagnostic delay by early identification of patients presenting characteristics common to CD and early initiation of CD-specific diagnostic pathways.

In recent years, clinical predictive model (CPM) trained on patients' electronic health records (EHR) have gained interest for prognostic or diagnostic tasks to identify new predictors and build clinical decision support systems [10]. In the context of IBD, to our knowledge CPMs have only been described for prognostic tasks (e.g., to predict disease complications or therapy response) [16]. In this work, we describe the extraction of an EHR-based CD incident cohort and matched controls from the Mount Sinai Health System (MSHS) and subsequent prediction of CD diagnosis using features derived from structured EHR and clinical notes.

2 Methods

2.1 Data and Study Population

The data used in this study stems from the Mount Sinai Data Warehouse (MSDW) which contains structured EHR data and unstructured clinical notes for approximately 10.5 million patients in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We included data from November 1st, 2011, to December 31st, 2021, in our analyses.

2.2 Phenotyping Algorithm

We applied EHR-based phenotyping to identify CD incident cases, the date of their first CD diagnosis, and matched controls from MSDW (Figure 1). Our phenotyping algorithm was developed by iterative investigations of the raw data

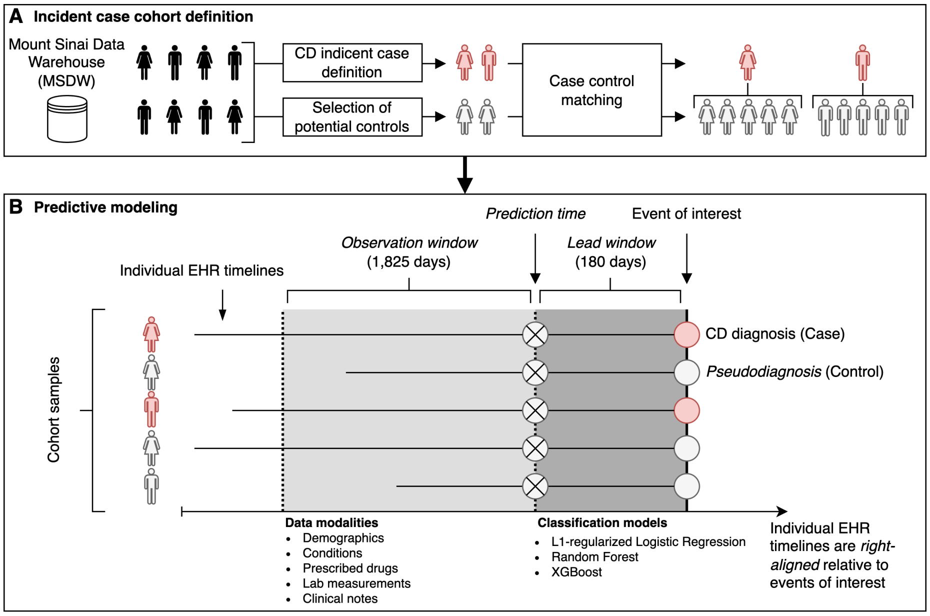
A Incident case cohort definition
Mount Sinai Data Warehouse (MSDW)
CD indicent case definition
Selection of potential controls
Case control matching
B Predictive modeling
Prediction time
Event of interest
Observation window (1,825 days)
Lead window (180 days)
Individual EHR timelines
Cohort samples
CD diagnosis (Case)
Pseudodiagnosis (Control)
Data modalities
Demographics
Conditions
Prescribed drugs
Lab measurements
Clinical notes
Classification models
L1-regularized Logistic Regression
Random Forest
XGBoost
Individual EHR timelines are right-aligned relative to events of interest

Fig. 1: Setup of Crohn's disease incident case cohort definition (A) and predictive modeling (B). Figure partially adapted from Lauritsen et. al [14]

contained in primary clinical information systems of randomly selected cases and controls.

We defined CD cases as patients with at least one IBD-specific medication prescription and two or more CD diagnosis codes coded on different days [11]. To select cases that had their incident diagnosis (index date) within MSHS (i.e., are not referral cases), healthcare encounters with non-CD diagnoses were required in the first and second year prior to the index date, and the first coded CD diagnosis within the context of an on-site encounter (excluding telehealth encounters). The former filtering step additionally ensured data availability within the observation time frame. We excluded patients with evidence of an existing IBD condition prior to the index date (i.e., specific medication, history of IBD in structured or text data). To exclude known CD patients based on unstructured clinical notes, we devised a list of strings indicative of prevalent IBD. To reduce the probability of including CD patients into the control group we excluded individuals that had an IBD condition or IBD-specific medication coded, or the presence of an IBD condition could be identified from clinical notes at any time in their EHR.

Cases and controls were exactly matched in a 1:5 ratio on year of birth in five-year bins, sex (female/male), and healthcare utilization metrics as described by Castro et. al [6]. Healthcare utilization, i.e., first and last recorded database entry and the total number of entries, was estimated from records of medication prescriptions, observation entries, diagnoses, or procedures. These features were transformed into a uniform distribution by quantile transformation and grouped in eight bins for matching [6]. Each control was then assigned the exact index date of their matched case as a *pseudodiagnosis* date [4].

2.3 Risk Prediction Problem Framing

Our CD risk prediction approach is outlined according to the methodology delineated in Lauritsen et. al [14] (Figure 1). The EHR data of each individual of the cohort was *right-aligned* relative to the event of interest (index or *pseudodiagnosis* date). Taking into account the previously described CD diagnostic delay, the *prediction time* was set at 180 days (approximately 6 months) before the event of interest. Features from the EHR used for predictive modeling and cohort characterization were extracted from the *observation window* defined as 2,005 to 180 days before the event of interest, spanning approximately 5 years. The EHR data between *prediction time* and the event of interest was excluded from analysis (*lead window*).

2.4 Data Pre-processing and Feature Extraction

EHR data included in the *observation window* of each individual was extracted from MSDW. Depending on the data modality, different pre-processing methods were applied to aggregate the data from the *observation window*.

Structured data included drug prescriptions, condition codes, demographics and measurements. Drug prescriptions and condition codes were processed to Boolean features, indicating whether a drug or condition was recorded at any time in the *observation window* of the individual. Numeric measurements were aggregated by calculating the median, maximum, and minimum value measured during the *observation window*, additionally the absolute count per measurement type was used as feature. Missing values were imputed by median imputation. Any condition or drug coded in less than 0.1% or measurements coded in less than 5% of individuals were removed. The age in years of each individual was calculated at *prediction time*. Sex, race, smoking status were extracted from the corresponding data tables and represented as categorical features (i.e., female/male, white/non-white/unknown and never smoker/smoker/ex-smoker/unknown). The unstructured clinical notes were cleaned by deleting duplicate notes and texts shorter than three words and aggregated per individual. Notes were encoded by term frequency-inverse document frequency (TF-IDF) with stop-word removal.

We used structured features only or in combination with text features as inputs. Boruta feature selection was applied on each dataset [13].

2.5 Predictive Modeling and Evaluation

We applied three different classification models: XGBoost [7], Random Forest (RF) [5] and Logistic Regression (with L1- or L2-regularization). The data was split stratified by class into a training set (70%) for model building and a test set (30%) for performance evaluation. The training set was used for the selection of model hyperparameters by 5-fold cross validation using Bayesian hyperparameter optimization [3]. Model performance was compared based on the area under the receiver operating characteristic (AUROC), area under the precision recall curve (AUPRC), F1, and accuracy. For model explainability, we used the SHapley Additive exPlanations (SHAP) method [15].

3 Results

To extract CD incident cases within the MSHS, multiple criteria including coded conditions, prescribed medication, and healthcare utilization had to be fulfilled. Using the phenotyping algorithm we identified 7,582 likely CD cases with at least three CD diagnosis codes on different days and prescription of IBD-specific medication. To exclude potential referral cases this number was reduced to 249 incident cases by filtering based on MSHS utilization prior to the index date and due to the requirement of the first coded CD diagnosis being recorded at an on-site visit. Cases were matched to controls based on age, sex and healthcare utilization, if available in a 1:5 ratio (Table 1). For the case cohort, we observe two peaks in the age distribution of the first CD diagnosis, in their second and fourth decade of life, consistent with known epidemiological patterns [18].

Table 1: Demographic and smoking information of the 249 CD incident cases and 1,242 controls included in the study. Hypothesis testing: Kruskal-Wallis or chi-squared test with Bonferroni correction

	Controls	**Cases**	**p-value**
n	1,242	249	
Age at prediction in years median (Q1,Q3)	38.0 (25.0,60.0)	38.0 (25.0,60.0)	0.958
Sex = female (%)	59.9	59.8	1.000
Race (%)			<0.001
White	45.4	70.7	
Black or African American	15.0	10.8	
Asian	4.0	0.3	
Native Hawaiian or Other Pacific Islander	0.3	0.0	
Other	10.4	23.0	
Unknown	12.4	5.2	
Smoking (%)			0.561
Smoker	5.6	7.6	
Ex-Smoker	14.5	15.3	
Never	49.4	46.2	
Unknown	30.5	30.9	

We compared condition prevalence in cases and controls. GI conditions, such as anal fistula or abnormal stool findings, were significantly overrepresented in cases prior to the first coded CD diagnosis (Table 2). No conditions were underrepresented in cases compared to controls. Blackwell et al. developed an IBD symptoms list, which groups GI symptoms into three categories: rectal bleeding, diarrhea, and abdominal and perianal pain [4]. The prevalence of all symptom groups were significantly overrepresented in cases. In total, 33% of cases had a coded IBD symptom in comparison with 12% of the control cohort 180 days before their first coded CD diagnosis. For the 144 cases with GI symptom coded any time before first CD diagnosis, the mean time span between these two codes, a potential estimation of the diagnostic delay, was 23.5 months (standard deviation (SD)=28.8, median=11.7).

Table 2: Conditions with significant overrepresentation in the CD incident case cohort in comparison to controls. Hypothesis testing: Fisher's Exact Test with false discovery rate (FDR) adjustment (q-value). OR, odds ratio.

	Prevalence (%)			
Conditions	**Case**	**Control**	**OR**	**q-value**
Anal fistula	2.81	0.08	35.90	0.010
Stool finding	2.81	0.16	17.93	0.022
Hemorrhage of rectum and anus	3.21	0.24	13.71	0.017
Rectal hemorrhage	5.22	0.40	13.63	<0.001
Anal fissure	4.02	0.40	10.35	0.010
Irritable bowel syndrome	6.02	0.64	9.89	<0.001
Generalized abdominal pain	4.82	0.81	6.24	0.017
Diarrhea	11.65	2.33	5.51	<0.001
Nonspecific abdominal pain	7.23	1.77	4.32	0.010
Abdominal pain	6.83	1.69	4.26	0.014
Gastroesophageal reflux disease	11.65	4.67	2.69	0.022
Grouped symptoms [4]	**Case**	**Control**	**OR**	**q-value**
Rectal bleeding	10.84	1.69	7.07	<0.001
Diarrhea	14.06	3.86	4.07	<0.001
Abdominal and perianal pain	20.88	8.21	2.95	<0.001
Any gastrointestinal symptom	32.93	12.08	3.57	<0.001

Using EHR data from the *observation window* of each individual, we built machine learning (ML) models to predict the risk of a CD diagnosis code 180 days after the *prediction time* (Table 3). In total 1,637 features from structured EHR were used as model input (901 conditions, 660 drugs, 64 measurement, and 12 demographic features). From 22,204 clinical notes (mean count per individual: cases 12.9, controls 15.3), depending on the optimal TF-IDF hyperparameter combination, between 12,358 and 1,355,059 text features were included. The EHRs of only 68% of controls and 73% of cases include at least one clinical note in MSDW2 during the *observation window*. Boruta Feature selection reduced the dataset consiting of only structured EHR data to 1,187 features, and the combined dataset to 3,690 features. XGBoost trained with only structured EHR data achieved highest AUROC, AUPRC, and F1. All classification models had lower performance in terms of AUROC if trained on structured EHR data and text.

Table 3: Model performance comparison of different machine learning algorithms data input. LR, logistic regression

Model	**Data**	**AUROC**	**AUPRC**	**F1-Macro**	**Accuracy**
XGBoost	Structured EHR	**0.72**	**0.44**	**0.65**	0.80
Random Forest		0.69	0.39	0.62	**0.83**
LR (L1-regularized)		0.69	0.34	0.60	0.69
XGBoost	Structured and text EHR	0.70	0.39	0.62	0.77
Random Forest		0.65	0.28	0.58	0.75
LR (L2-regularized)		0.68	0.35	0.59	0.70

To explain the predictions made by the best performing model based on AUROC, we analyzed the predictions using SHAP (Figure 2). Coded *White* race had the largest impact on model predictions, increasing the likelihood of case classification. The majority of features with high prediction influence were numerical measurements, comprised of anemia-related, electrolyte and blood count lab values. The coding of specific conditions (diarrhea and gastroesophageal reflux disease) and antibiotic and anti-inflammatory drugs (ciprofloxacin and prednisone) increased the likelihood to classify individuals as at risk for CD diagnosis.

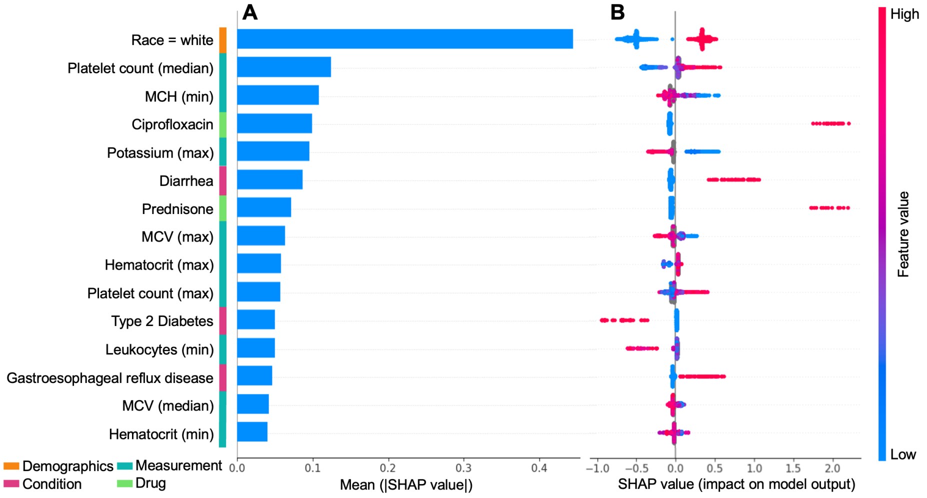

Fig. 2: SHAP values of the XGBoost model (structured EHR only). Mean SHAP value (A) and the SHAP value of individual predictions (B) of the 15 most informative features

4 Discussion

4.1 Clinical and Technical Significance

Using the GI symptom groups developed by Blackwell et. al [4], we can confirm the increased symptom prevalence for abdominal and perianal pain, diarrhea and rectal bleeding in our cohort, at least 180 days prior to the first CD diagnosis. Compared to Blackwell et al., we report higher prevalence of these symptoms for both cases and controls, potentially due to the longer 5 year *observation window* in our study compared to the 12 months. In our best performing CPM, only few of these overrepresented conditions are amongst the 20 most important features.

CPMs with features from both structured EHR and clinical notes often perform better than models that are based on only one of the two data modalities [17]. To enable the identification of yet unknown/unexpected features prior to CD diagnosis, we encoded the clinical text by TF-IDF, an unsupervised method to weight text terms by their appearance frequency in single documents and the whole corpus. The high dimensionality of our input feature matrix when adding TF-IDF-encoded vectors as well as the absence of clinical text in 31% of patients may explain the reduced performance compared to the structured EHR data alone. The advantage of using text information in this study shown for the phenotyping of CD incident cases: Since many CD patients are referred to the MSHS as tertiary care center and are not initially diagnosed on site, we stringently filtered out CD cases with previously diagnosed disease which was not captured sufficiently in the structured data.

4.2 Limitations and Future Work

While our study shows promising results, we acknowledge a number of limitations. First, only the first part of our phenotyping algorithm, the identification of CD patients, has shown to have high sensitivity and specificity [11]. A validation of the CD incident case cohort is further required. With a larger cohort and external validation of our results in a second hospital system, we will be more confident in the generalizability of our results.

Further limitations apply to the nature of the data that we use for our study: clinical research using EHR data is challenging, amongst others due to data quality issues, for instance caused by a data collection bias and missingness in the data, or with regards to accuracy of a patients' ethnicity [1,2]. EHR data recorded during the *observation window* of 1825 days was aggregated in this study. Using advanced prediction models that incorporate temporal information, e.g., recurrent neural networks, or optimizing aggregation based on different time frames might further improve model performance.

We also recognize that the framing of our study would be more applicable to the clinical use case by setting up the prediction model with *left-aligned* patient data, thus having a common prediction time point on a common event across controls and cases [14]. In CD this time point could be defined by the first presentation of GI symptoms. We did not pursue this strategy since only 33 % of cases had coded GI symptoms in their structured EHR, resulting in a very small study cohort. To further investigate the magnitude of a potential acceleration of CD diagnosis, predictive modeling with varying lengths of *lead windows* could be explored. Comparing the feature importance between different prediction time points might reveal early identifiable risk features of prospective CD cases.

To further improve the discriminative ability of our models, we are working with the clinical notes in a more supervised manner by extracting and aggregating specific symptoms and conditions. In addition, the aggregation of terms might reduce high dimensionality of our input data. This could be conducted in a supervised manner (e.g., applying the groups described by Blackwell et al. [4]) or by linking terms to biomedical concepts on which hierarchical aggregation could be performed.

5 Conclusion

To our knowledge, this is the first study to describe an EHR-based phenotyping algorithm to identify CD incident cases as well as a diagnostic CPM to predict CD cases prior to their clinical diagnosis. With our best performing ML algorithm, XGBoost, we achieved an AUROC of 0.72 and AUPRC of 0.44, demonstrating the feasibility of this prediction task, though clinical validation of our results is still pending. The high overrepresentation of GI symptoms more than six months prior to the actual diagnosis in our cohort at the MSHS underpins the need to reduce CD diagnostic delay, even at a tertiary care center with a focus on IBD.

6 Acknowledgements

This work is supported in part through the use of the research platform AI-Ready Mount Sinai (AIR.MS), and through the MSDW resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai. The research leading to these results has received funding from the Horizon 2020 Programme of the European Commission under Grant Agreement No. 826117 and by the Joachim-Herz foundation (to SI).

References

1. Beaulieu-Jones, B.K., Lavage, D.R., Snyder, J.W., Moore, J.H., Pendergrass, S.A., et al.: Characterizing and managing missing structured data in electronic health records: Data analysis. JMIR Med. Inf. **6**(1), e11 (2018)
2. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., et al.: Toward a fine-scale population health monitoring system. Cell **184**(8), 2068–2083 (2021)
3. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on Machine Learning. pp. 115–123. PMLR (2013)

4. Blackwell, J., Saxena, S., Jayasooriya, N., Bottle, A., Petersen, I., Hotopf, M., et al.: Prevalence and duration of gastrointestinal symptoms before diagnosis of inflammatory bowel disease and predictors of timely specialist review: A population-based study. JCC **15**(2), 203–211 (2021)
5. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
6. Castro, V.M., Apperson, W.K., Gainer, V.S., Ananthakrishnan, A.N., Goodson, A.P., Wang, T.D., et al.: Evaluation of matched control algorithms in EHR-based phenotyping studies: A case study of inflammatory bowel disease comorbidities. J. Biomed. Inform. **52**, 105–111 (2014)
7. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)
8. Danese, S., Fiorino, G., Mary, J.Y., Lakatos, P.L., D'Haens, G., Moja, L., et al.: Development of red flags index for early referral of adults with symptoms and signs suggestive of Crohn's disease: An IOIBD initiative. JCC **9**(8), 601–606 (2015)
9. Fiorino, G., Bonovas, S., Gilardi, D., Di Sabatino, A., Allocca, M., Furfaro, F., et al.: Validation of the red flags index for early diagnosis of Crohn's disease: A prospective observational IG-IBD study among general practitioners. JCC **14**(12), 1777–1779 (2020)
10. Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.P.A.: Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. J. Am. Med. Inf. Assoc. **24**(1), 198–208 (2017)
11. Ibing, S., Cho, J.H., Böttinger, E.P., Ungaro, R.C.: Second line biologic therapy following tumor necrosis factor antagonist failure: A real world propensity score weighted analysis. CGH (in press) (2023)
12. Jayasooriya, N., Baillie, S., Blackwell, J., Bottle, A., Petersen, I., Creese, H., et. al: Systematic review with meta-analysis: Time to diagnosis and the impact of delayed diagnosis on clinical outcomes in inflammatory bowel disease. Aliment. Pharmacol. Ther. **57**(6), 635–652 (2023)
13. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Stat. Soft. **36**(11), 1–13 (2010)
14. Lauritsen, S.M., Thiesson, B., Jørgensen, M.J., Riis, A.H., Espelund, U.S., Weile, J.B., et al.: The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. npj Digit. Med. **4**(1), 1–12 (2021)
15. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020)
16. Nguyen, N.H., Picetti, D., Dulai, P.S., Jairath, V., Sandborn, W.J., Ohno-Machado, L., et al.: Machine learning-based prediction models for diagnosis and prognosis in inflammatory bowel diseases: A systematic review. JCC **16**(3), 398–413 (2022)
17. Seinen, T.M., Fridgeirsson, E.A., Ioannou, S., Jeannetot, D., John, L.H., Kors, J.A., et al.: Use of unstructured text in prognostic clinical prediction models: a systematic review. J. Am. Med. Inf. Assoc. **29**(7), 1292–1302 (2022)
18. Torres, J., Mehandru, S., Colombel, J.F., Peyrin-Biroulet, L.: Crohn's disease. Lancet **389**(10080), 1741–1755 (2017)
19. Ungaro, R., Mehandru, S., Allen, P.B., Peyrin-Biroulet, L., Colombel, J.F.: Ulcerative colitis. Lancet **389**(10080), 1756–1770 (2017)