Diagnostic Accuracy Study

Medicine®

OPEN

A novel surgical predictive model for Chinese Crohn's disease patients

Yuan Dong, MD[a], Li Xu, MD[b], Yihong Fan, MD[c], Ping Xiang, MD[d], Xuning Gao, MD[d], Yong Chen, MD[e], Wenyu Zhang, PhD[e], Qiongxiang Ge, PhD[b,*]

Abstract

Due to the complexity of Crohn's disease (CD), it is difficult to predict disease course with a single stratification factor or biomarker. A logistic regression (LR) model has been proposed by Guizzetti et al to stratify patients with CD-related surgical risk, which could help decision-making on disease treatment. However, there are no reports on relevant studies on Chinese population. The aim of the study is to present and validate a novel surgical predictive model to facilitate therapeutic decision-making for Chinese CD patients. Data was extracted from retrospective full-mode electronic medical records, which contained 239 CD patients and 1524 instances. Two sub-datasets were generated according to different attribute selection strategies, both of which were split into training and testing sets randomly. The imbalanced data in the training sets was addressed by synthetic minority over-sampling technique (SMOTE) algorithm before model development. Seven predictive models were employed using 5 popular machine learning algorithms: random forest (RF), LR, support vector machine (SVM), decision tree (DT) and artificial neural networks (ANN). The performance of each model was evaluated by accuracy, precision, F1-score, true negative (TN) rate, and the area under the receiver operating characteristic curve (AuROC). The result revealed that RF outperformed all other baseline models on both sub-datasets. The 10 leading risk factors for CD-related surgery returned from RF for attribute ranking were changes of radiology, presence of a fistula, presence of an abscess, no infliximab use, enteroscopy findings, C-reactive protein, abdominal pain, white blood cells, erythrocyte sedimentation rate and platelet count. The proposed machine learning model can accurately predict the risk of surgical intervention in Chinese CD patients, which could be used to tailor and modify the treatment strategies for CD patients in clinical practice.

Abbreviations: ANN = artificial neural networks, AuROC = area under the receiver operating characteristic curve, CAD = computer-aided diagnosis, CD = Crohn's disease, DT = decision tree, IBD = inflammatory bowel disease, LR = logistic regression, RF = random forest, SMOTE = synthetic minority over-sampling technique, SVM = support vector machine, TN = true negative.

Keywords: Crohn's disease, machine learning, surgical predictive model

1. Introduction

Crohn's disease (CD) is a chronic inflammatory bowel disease (IBD) that can affect any part of the gastrointestinal tract, triggering persistent transmural inflammation, followed by structural intestinal damage and intestinal complications, such as strictures, fistulas, and abscesses that often require surgery.[1] Even though surgery is not curative, it plays an indispensable role in the treatment of CD. Approximately 70% to 90% of patients require a type of surgery at some point in their lives after being diagnosed with CD.[2] Steroid dependence, young age at diagnosis, current smoking, small bowel involvement, perianal disease at the time of diagnosis, and penetrating or stricturing disease behavior are the usual predictors of a disabling disease course, which in turn is associated with greater surgical requirements.[3–5] However, the prevalence of at least 2 of these factors is often very high, which makes it difficult to guide treatment in clinical practice.[6] More importantly, postoperative mortality is affected by the timing of surgery because CD patients who require emergency surgery experience more complications than patients who undergo selective surgery.[7] There is no question that well-timed surgical interventions can decrease the odds of emergency surgery and stoma, as well as lead to long-term disease relief while improving quality of life.[8] Early identification of patients who subsequently undergo surgery can initiate early intensive treatment to alter the natural course of the disease or at least to optimize perioperative management so that patients are better able to tolerate surgery and reduce postoperative complications. Therefore, it would be extremely useful to establish a model with multidimensional attributes that could predict the risk of surgery.

With the help of information technology, computer-aided diagnosis (CAD) has brought about tremendous changes in

Editor: Goran Augustin.

National Natural Science Foundation of China (No. 81473506).

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

The authors have no conflicts of interests to disclose.

[a] Department of Pathology, [b] Department of Anorectal Surgery, [c] Department of Gastroenterology, [d] Department of Radiology, The First Affiliated Hospital of Zhejiang Chinese Medical University, Zhejiang Provincial Hospital of TCM, Zhejiang International Exchange Center of Clinical TCM, [e] School of Information, Zhejiang University of Finance and Economics, Hangzhou 310018, China.

[*] Correspondence: Qiongxiang Ge, Department of Anorectal Surgery, The First Affiliated Hospital of Zhejiang Chinese Medical University, Zhejiang Provincial Hospital of TCM, Zhejiang International Exchange Center of Clinical TCM, 54 Youdian Road, Hangzhou 310006, China (e-mail: geqiongxiang@163.com).

How to cite this article: Dong Y, Xu L, Fan Y, Xiang P, Gao X, Chen Y, Zhang W, Ge Q. A novel surgical predictive model for Chinese Crohn's disease patients. Medicine 2019;98:46(e17510).

Received: 15 February 2019 / Received in final form: 20 August 2019 / Accepted: 12 September 2019

http://dx.doi.org/10.1097/MD.0000000000017510

clinical decision-making.[9] In recent years, machine learning (ML) and data mining models have been well applied in the clinical field, and various high-performance models are expected to help clinicians predict disease progression and adjust treatment over time. However, there is little reference on the application of ML models for CD-related decision-making, except for a risk predictive model for CD-related surgery based on logic regression (LR), which involves 10 selected baseline predictor attributes, developed and validated by Guizzetti et al.[10] It is believed that once externally validated, their predictive model could be used to guide CD management.

The extrapolated disease incidence and prevalence rates of CD in China are $0.848/10^5$ and $2.29/10^5$ person/year, respectively, showing a rapid upward trend, which will cause enormous medical and economic burdens.[11] However, due to the differences of genetic variation, epidemiology, and clinical phenotypes between East Asians and whites,[12] it is not appropriate to apply Guizzetti et al's model[10] on Chinese CD patients directly. In addition, there are no reports on relevant studies on the Chinese population. In light of these facts, we explored the ML method for developing a novel predictive model of CD-related surgery for Chinese patients.

In this study, our goals were to

(1) apply and evaluate the suitability of Guizzetti et al's predictive model[10] and corresponding predictive attributes on Chinese CD patients; and
(2) present and validate a novel surgical predictive model to facilitate therapeutic decisions for Chinese CD patients, as well as recommend the right surgical time in daily clinical practice.

Guizzetti et al's method[10] based on LR and the corresponding predictor attributes were evaluated first on our dataset. The random forest (RF) algorithm for prediction was also applied to the same data with the same attributes for comparison of predictive performance. It was found that RF predicted the CD-related surgery more accurately than LR. Then, RF was applied to the original dataset for prediction again with the predictor attributes reselected optimally. We compared it with 4 other widely used classification algorithms—namely, LR, support vector machines (SVM), decision trees (DT), and artificial neural networks (ANN)—to choose the best model for clinical application. The experiment showed that RF outperformed all other baseline algorithms in prediction.

2. Patients and methods

2.1. Patient population

The study sample included all CD patients who were hospitalized at or clinically visited the First Affiliated Hospital of Zhejiang Chinese Medicine University from January 1, 2010 to July 31, 2018. The patients' sample included 239 patient individuals corresponding to 1524 instances. An instance was defined as each clinical event, including each inpatient or outpatient record with an objective assessment of inflammation, either C-reactive protein (CRP) or fecal calprotectin (FCP). Duplicated instances were deleted for the same patients if their returned hospitalization or visit were within 30 days. Among the 239 CD patients, 141 (59%) were male, and the gender ratio (male/female) was 1.43. The mean age was 35.36 years and the age range was 3 to 80 years.

In this study, 92 patient individuals and 458 instances were excluded by applying 4 exclusion criteria:

(1) patients who did not receive CD-related treatment during hospitalization;
(2) patients without evaluation of inflammation, whether CRP or FCP;
(3) patients with less than 2 records;
(4) instances with more than 50% missing data (shown in Fig. 1).

Of the 92 excluded patients, 7 had only one record without any indicators of inflammation assessment (CRP or fecal calprotectin) or any CD-related treatment. Fourteen out of 92 excluded patients, for whom more than 50% data are missing, did not receive CD-related treatment. This study was approved by the First Affiliated Hospital of Zhejiang Chinese Medicine University Research Ethics Committee [2014-K197]. All patients consented for their data to be analyzed.

2.2. Data extraction

Data were manually extracted from retrospective full-mode electronic medical records (EMR) captured during patient admissions, outpatient visits, or emergency visits, which was a time-consuming and labor-intensive part of this study. The 131 data items were divided into 6 categories: demographics, clinical presentations, laboratory tests, auxiliary examinations, medical orders, and nursing records (shows in Table 1).

2.3. Definitions

Patients were diagnosed with CD based on a combination of standard criteria, including clinical symptoms, endoscopy, histopathology, radiology, and/or biochemical investigations.[13] Stricturing disease was defined as luminal narrowing, which was impossible or difficult to pass through an adult endoscope, or obstructive symptoms without penetrating disease. Penetrating disease was defined as the presence of intra-abdominal fistulas, inflammatory masses, or abscesses. CD-related surgery was defined as any surgical procedure for CD, including abdominal surgery (ileal resection, ileocecal resection, colectomy, proctocolectomy, proctectomy, stricturoplasty, segmental resection and anastomosis with or without diversion, enterourinary fistulas repair, and ostomy formation and repair), and perianal surgery (incision and drainage of abscess, debridement, seton placement, fistulotomy, fistulectomy, ligation of intersphincteric fistula tract advancement flap closures, and rectovaginal fistula repair).

2.4. Data preprocessing

Two of the biggest challenges with the EMR data were the missing values and low-quality data, therefore, the data preprocessing was indispensable for improving predictive performance. The main data preprocessing in this study was depicted as the following parts:

(1) Classifying attributes: The main attributes, their data type and domain values are shown in Table 2. The last attribute, *Surgery*, was the classification label of the original dataset and took 2 categorical integers: 1 for CD-related surgery and 0 for no surgery. Changes of radiology from different types of imaging were integrated into five radiological characteristics, as shown in Table 2, which were blind reviewed by 2 radiologists.

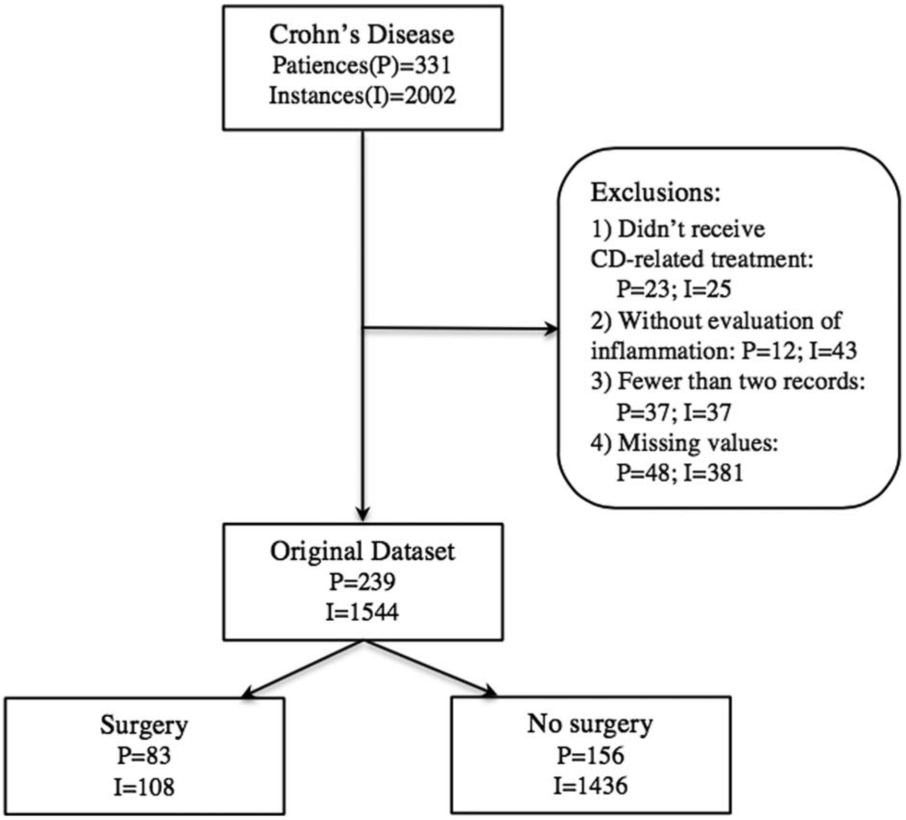
Crohn's Disease
Patiences(P)=331
Instances(I)=2002
Exclusions:
1) Didn't receive CD-related treatment: P=23; I=25
2) Without evaluation of inflammation: P=12; I=43
3) Fewer than two records: P=37; I=37
4) Missing values: P=48; I=381
Original Dataset
P=239
I=1544
Surgery
P=83
I=108
No surgery
P=156
I=1436

Figure 1. STARD diagram for original dataset. I=instance, P=patients.

(2) Reducing invalid attributes: Private information (such as *ID card number*, *telephone number*, *address*, and so on), and redundant attributes (such as *admission time*, *vital signs*, *cost*, *allergen*, *nursing records*, and so on) were deleted to reduce their side effects on the predictive accuracy and efficiency of the classification model. If an attribute had more than 50% missing values, such as *FCP*, *C. difficile*, *ASCA*, *pANCA*, *ANCA*, *anti-OmpC*, *anti-CBir1*, *parenteral manifestations*, *oral ulcerations*, *anxiety status*, and so on, it was deleted from the original dataset. *Smoking status* was also deleted from the attributes, by considering that the Confucian culture in China has caused young people and women to conceal smoking habits,[14,15] which means the related data abstracted from EMR may be inaccurate.

(3) Filling missing data: No instances had missing classification label. Missing values were filled based on the mean value of data from the same attribute of the same patient. If a patient had fewer than 3 records, we imputed the mean value of that attribute from all patients.

(4) Sub-dataset generation and attributes reselection: 2 sub-datasets were generated according to different attribute selection strategies:

- Sub-dataset 1 contained 10 attributes according to Guizzetti et al's surgical predictive model,[10] except for the use of the *Crohn's Disease Activity Index (CDAI)* instead of the *Harvey-Bradshaw Index (HBI)* because *HBI* (as a simpler version of *CDAI*) was not available in our original dataset, as shown in Table 2;

Table 1

Metadata extracted from EMR.

Demographics	Gender, age, ID card number, education, address, telephone number, occupation, disease duration, family history, smoking status, prior surgery history, time of admission, cost, etc.
Clinical presentations	Abdominal mass, abdominal pain, stool frequency, Bristol classification, hemorrhage, mucus in stool, parenteral manifestations, oral ulcerations, perianal lesions, rectal involved, fever, anxiety status, etc.
Laboratory tests	FBC, U&E, LFT, coagulation test, CRP, ESR, FCP, C. difficile, ASCA, pANCA, ANCA, anti-OmpC, anti-CBir1, etc.
Auxiliary examinations	Endoscopy: colonoscopy, upper GI endoscopy, ileoscopy, capsule endoscopy Cross-sectional imaging: MRI and CT enterography, pelvic MRI
Medical orders	Medication exposure: immunosuppressives, 5-aminosalicylate, corticosteroids, infliximab, probiotics, enteral nutrition, antibiotics, herb medicine, etc. Procedure: surgery, balloon dilation, enteral feeding (tube feeding), endoscopic hemostasis, etc.
Nursing records	Vital signs, BMI, VAS, etc.

ANCA=antineutrophil cytoplasmic autoantibody, anti-OmpC=Escherichia coli outer membrane porin C, ASCA=anti-saccharomyces cerevisiae antibody, BMI=body mass index, CRP=C-reactive protein, EMR=electronic medical records, ESR=erythrocyte sedimentation rate, FBC=full blood count, FCP=fecal calprotectin, LFT=liver function test, pANCA=perinuclear antineutrophil cytoplasmic antibody, U&E=urea and electrolytes, VAS=Visual Analogue Scale; Vital signs=temperature, blood pressure, pulse, and breathing rate.

Table 2

The main attributes information of both sub-datasets.

Attribute name	Data type	Domain	Missing value	Sub-dataset
Gender	Binary	0 = male; 1 = female	0	1
CDAI	Numerical	1-387	788	1
5-Aminosalicylate use	Binary	0 = none; 1 = used	0	1
Age (in years)	Numerical	3-80	0	1, 2

Stool frequency	Nominal	–4 = stop pass gas;
		–3 = stop pass stool;
		–2 = 3 d/time;
		–1 = 2 d/time;
		Number[*] = times of defecation each day
			48	1, 2

Abdominal mass	Binary	0 = none; 1 = presented	0	1, 2
Presence of a fistula	Binary	0 = none; 1 = presented	0	1, 2
Presence of an abscess	Binary	0 = none; 1 = presented	0	1, 2

Disease location	Nominal	0 = none;
		1 = small bowel only[†];
		2 = colon only[‡];
		3 = colon + small bowel;
		4 = upper gastrointestinal tract
			1	1, 2

Immunosuppressives use	Binary	0 = none; 1 = used	0	1, 2

Pain	Nominal	0 = none;
		1 = intermittent dull pain;
		2 = persistent cramping pain;
		3 = persistent cramping pain with vomiting
			16	2

Mucus in stool	Nominal	0 = none; 1 = mucus	16	2
Disease duration (in months)	Numerical	0-266	8	2
FOBT	Numerical	0–4	75	2
WBC ($\times 10^9$/L)	Numerical	0.3–29.8	52	2
RBC ($\times 10^{12}$/L)	Numerical	1.67–6.56	54	2
PLT ($\times 10^9$/L)	Numerical	20-768	52	2
CRP (mg/L)	Numerical	0–257.7	57	2
ESR (mm/h)	Numerical	1–91	364	2
TT (seconds)	Numerical	1.6–33.2	142	2
PT (seconds)	Numerical	3.54–34.1	142	2
FIB (g/L)	Numerical	0.82-17	143	2
APTT (seconds)	Numerical	0.16–50.4	143	2
BUN (mmol/L)	Numerical	1.2–226	162	2
AST (g/L)	Numerical	2–248	147	2
ALB (g/L)	Numerical	4.5–87	141	2
GGT (U/L)	Numerical	1–652	147	2
BMI	Numerical	12.4–29.12	550	2

Enteroscopy findings	Nominal	0 = normal;
		1 = ulcers;
		2 = polyps;
		3 = ulcers + polyps;
		4 = stricture;
		5 = fistulous orifices;
		6 = bleeding
			133	2

Changes of radiology	Nominal	0 = normal;
		1 = inflammatory;
		2 = fibrotic strictures;
		3 = fistulas;
		4 = abscesses
			122	2

No infliximab use	Binary	0 = no; 1 = yes	0	2
Current cortisone use	Binary	0 = none;1 = used	0	2
History of appendectomy	Binary	0 = no; 1 = yes	0	2
Surgery	Binary	0 = no surgery; 1 = surgery	0	1, 2

ALB = serum albumin, APTT = activated partial thromboplastin time, AST = aspartate transaminase, BMI = body mass index, BUN = blood urea nitrogen, CRP = C-reactive protein, ESR = erythrocyte sedimentation rate, FIB = fibrinogen, FOBT = fecal occult blood test, GGT = gamma-glutamyl transpeptidase, PLT = platelet count, PT = prothrombin time, RBC = red blood cell count, TT = thrombin time, WBC = white blood cell count. If defecation is more than once a day, "Number" = times of defecation each day. small bowel: from jejunum to the ileocecal valve; colon: from the ileocecal valve to the anal verge.

[*] If defecation is more than once a day, "Number" = times of defecation each day.

[†] small bowel: from jejunum to the ileocecal valve.

[‡] colon: from the ileocecal valve to the anal verge.[32]

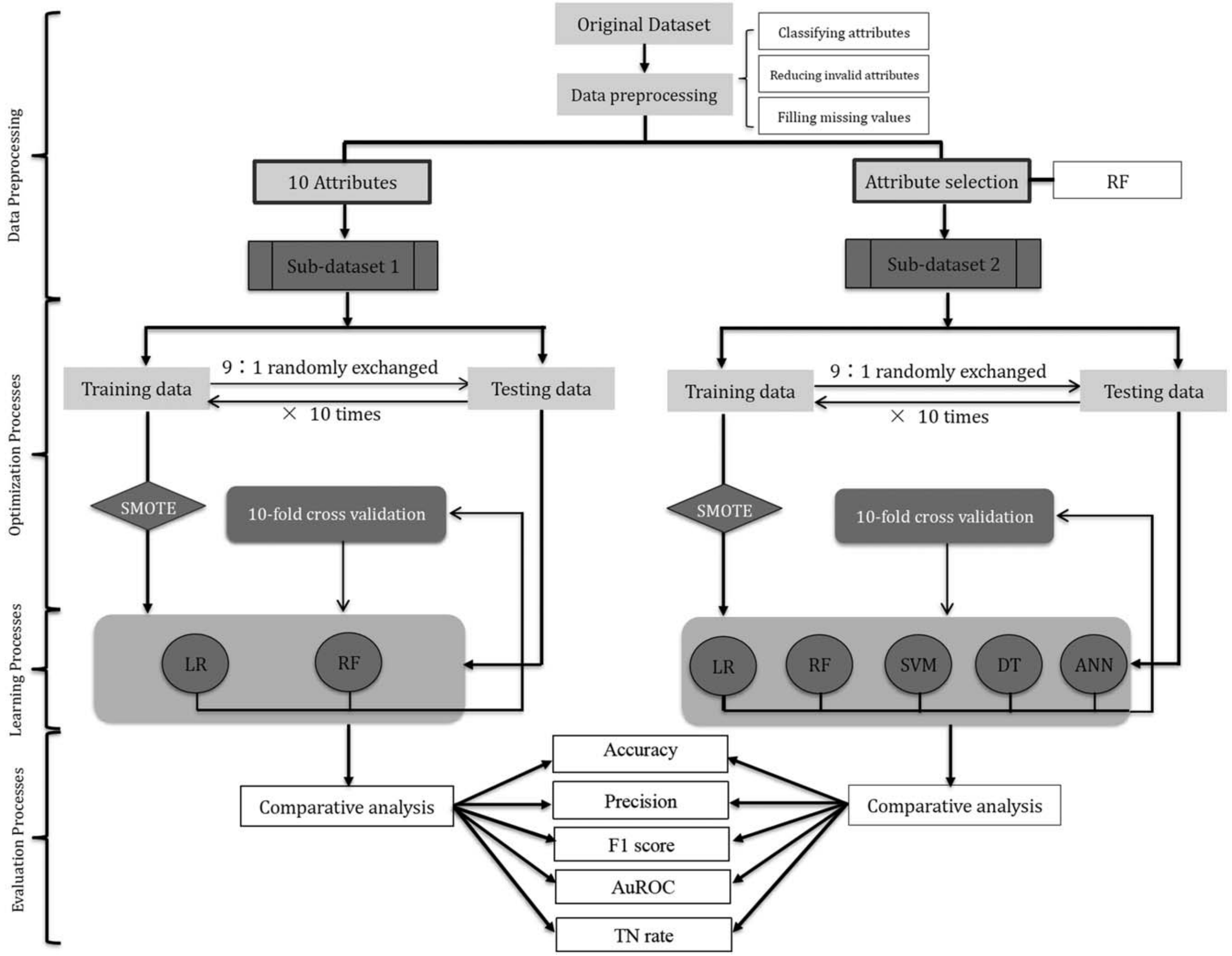

Figure 2. The flowchart of experiments. The 10 attributes in "Sub-dataset 1" included: age, gender, disease location, HBI, stool frequency, immunosuppressive use, 5-aminosalicylate use, presence of a fistula, presence of an abscess, and presence of an abdominal mass; the 30 attributes in "Sub-dataset 2" were selected by random forest (RF). Both sub-datasets were divided into the training set and the testing set with the proportion of 9:1. In learning processes: applying and validating logic regression (LR) and RF models on "Sub-dataset 1" with the synthetic minority over-sampling technique algorithm (SMOTE) on the left side, while applying and validating RF, LR, decision tree (DT), support vector machine (SVM) and artificial neural networks (ANN) on "Sub-dataset 2" in the similar way on the right side. 10-Fold cross-validation was performed to train and test the generated predictive models. Five metrics were collected: accuracy, precision, F1 score, true negative (TN) rate, and area under the receiver operating characteristic curve (AuROC).

- Sub-dataset 2 incorporated predictor attributes that were reselected by RF for attribute ranking from the main attributes shown in Table 2.

2.5. Modeling

A series of experiments were conducted on 2 sub-datasets and were implemented with Python 3.6 programming language. The flowchart of experiments proposed in this study is shown in Figure 2. The 10 attributes in "Sub-dataset 1" included: age, gender, disease location, CDAI, stool frequency, immunosuppressive use, 5-aminosalicylate use, presence of a fistula, presence of an abscess, and presence of an abdominal mass; the 30 attributes in "Sub-dataset 2" were selected by random forest (RF). Both sub-datasets were divided into the training set and the testing set with the proportion of 9:1. In learning processes: applying and validating logic regression (LR) and RF models on "Sub-dataset 1" with the synthetic minority over-sampling technique algorithm (SMOTE) on the left side, while applying and validating RF, LR, decision tree (DT), support vector machine (SVM) and artificial neural networks (ANN) on "Sub-dataset 2" in the similar way on the right side. Ten-Fold cross-validation was performed to train and test the generated predictive models. Five metrics were collected: accuracy, precision, F1 score, true negative (TN) rate, and the area under the receiver operating characteristic curve (AuROC).

The modeling process is divided into 3 parts:

- Optimization processes: Dealing with a training set with a heavy imbalance ratio by applying the synthetic minority over-sampling technique (SMOTE) algorithm.[16]
- Learning processes: Applying and validating LR and RF models on sub-dataset 1 with the SMOTE algorithm, while applying and validating LR, RF, SVM, DT, and ANN on sub-dataset 2 in the similar way.
- Evaluation processes: Evaluating each model via accuracy, precision, F1 score, TN rate, and AuROC.

All of the data used in the experiment was divided into 2 parts randomly—the training set and the testing set—with the proportion of 9:1. There were 108 instances that had surgeries, resulting in 93% of the patients in class 0% and 7% in class 1. Therefore, balancing the training set by over-sampling the under-representational classes through the proven over-sampling technique SMOTE could help ensure the balance of data and ensure classification accuracy in the minority class.[16]

The LR classification algorithm was used to model sub-dataset 1 on the left side. The LR model was trained in the training set processed by the SMOTE algorithm and then used for predicting the surgery probability in the testing set. The prediction results were compared with a real CD-related surgery event to measure the performance of each model. RF model was generated in the same way for comparison of predictive performance. After that, LR and the other four ML algorithms were used to model sub-dataset 2 on the right side. The models were trained and tested in the similar way, as shown in the flowchart.

Ten-Fold cross-validation was performed to train and test the generated predictive models in order to reduce the effect of the random choice of the training and testing sets. Due to the wide variety of evaluation methods, no one method can replace the other completely. To evaluate the performance of each model, four metrics were used in this study: accuracy, precision, F1 score, and TN rate, which were defined as follows:[17,18]

- Accuracy = True positives /(True negatives + False positives)
- Precision = True positives /(True positives + False positives)
- F1 score = (2 × Precision × Recall)/(Precision + Recall)
- TN rate = True negatives /(False positives + False negatives)
- Recall = True positives /(True positives + False negatives)

For more comprehensive comparison, AuROC was also calculated. In this study true positives were classified as a surgery instance when the patient underwent surgery in practice; false positives were classified as a surgery instance while the patient did not have surgery in practice; true negatives were classified as a non-surgery instance when the patient did not have surgery in practice; false negatives were classified as a non-surgery instance while the patient underwent surgery in practice.

3. Result

3.1. Demographics

Table 3 lists the demographics and basic characteristics of patients. Of the 83 patients that produce 108 surgery-operated instances, 61 patients underwent 1 surgery, 19 underwent 2 surgeries, and 3 underwent three surgeries. Other basic characteristics of patients are shown in Table 3.

3.2. Attributes reselection by RF

Besides its usage for prediction, the RF algorithm could also be used for attribute ranking by calculating the importance of each attribute. For original data, 100 trees were built to identify which attributes contributed most to the predictive accuracy of the model. The attributes with the smallest classification weight were removed from the forest to save calculating time and computational load, and ensure good generalization. The top 30 contributing attributes, selected by RF algorithms, were listed in Figure 3 in descending order, which included 7 attributes used in sub-dataset 1, except for the *gender*, *CDAI*, and *5-aminosalicylate use*. *Changes of radiology*, *presence of a fistula*, *presence of an abscess*, *no infliximab use*, *enteroscopy findings*, *CRP*, *abdominal pain*, *WBC*, *ESR*, and *PLT* were the top 10 attributes related to surgical events.

Table 3

Demographic and basic characteristics of patients.

	Number	Percentage
Total	239	
Sex: M:F	141:98	59%:41%
Median age (years)	35.36	
Disease location		
Small bowel only	31	13.0%
Colon only	62	25.9%
Small + Colon bowel	136	56.9%
Upper gastrointestinal tract	10	4.2%
Disease behavior		
Inflammatory	134	56.1%
Stricturing	67	28.1%
Penetrating	38	15.8%
Perianal disease	112	45.5%
Surgery	108	
Abdominal surgery	42	38.9%
Perianal surgery	66	61.1%

3.3. Development and evaluation of different models on 2 sub-datasets

The performance of the predictive models on sub-dataset 1 is presented in Figure 4. All metrics represent the average over the 10-folds of the cross validation. The results revealed that RF predictive model performed better than LR model in terms of accuracy (93.11% vs 91.15%), precision (53.42% vs 44.81%), F1 score (0.6016 vs 0.5763), TN rate (95.08% vs 92.00%), and the AuROC (0.8926 vs 0.8809), which implies that RF is more suitable than LR to model the complex real-world data about Chinese CD patients.

The performance of the predictive models on sub-dataset 2 is presented in Table 4. The results revealed that the average F1 score of RF predictive model was improved from 0.6016 (in sub-dataset 1) to 0.7706 (in sub-dataset 2) due to attribute reselection by RF algorithm. The average F1 score of LR predictive model was also improved from 0.5763 (in sub-dataset 1) to 0.6308 (in sub-dataset 2) due to attribute reselection by RF algorithm. The average F1 score of the other three models on sub-dataset 2 were 0.7112 for DT, 0.6288 for SVM, and 0.5757 for ANN, respectively. This implies that the average F1 score of RF predictive model is the best among all 5 predictive models under comparison. The AuROCs were still excellent at 0.9864 in RF, 0.9538 in LR, 0.8809 in DT, 0.9497 in SVM, and 0.9059 in ANN, respectively (Fig. 5). Similarly, it was found that the average accuracy, precision and TN rate of RF predictive model are the best among all 5 predictive models, respectively. As a result, RF performed best in prediction on both sub-datasets.

4. Discussion

Due to the complexity of CD, it is impossible to identify a single stratification factor or biomarker for patient populations.[19] Since 2006, when Beaugerie et al[20] first pointed out that steroid-dependent young CD patients with perianal disease involved at diagnosis was a disabling form of CD, more and more risk factors

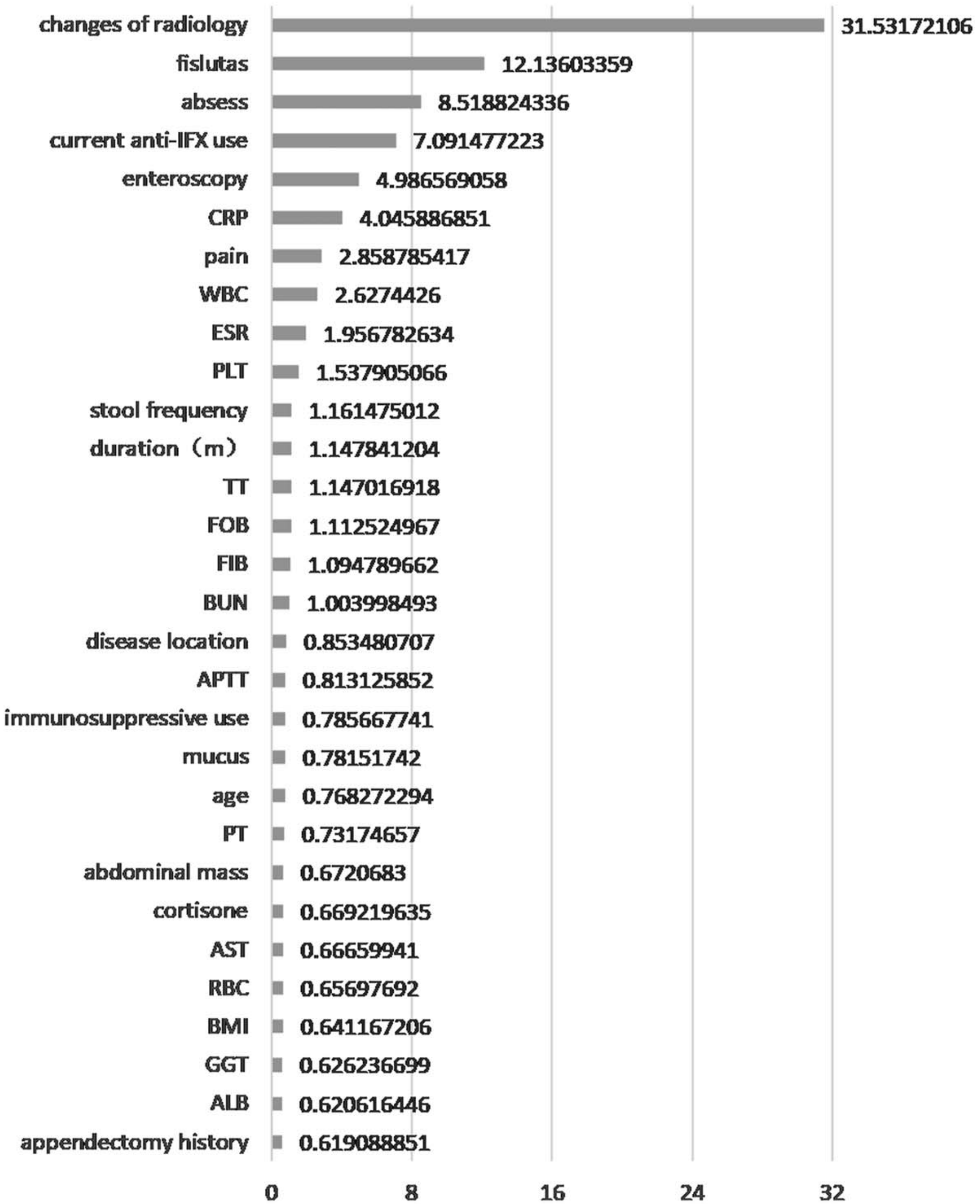

Figure 3. Top 30 attributes with the highest importance reselected by RF. The importance of each attribute was quantitatively assessed by the number of occurrences of each one in the 100 runs by RF algorithms. The top 30 contributings, selected by RF were listed on the *y*–axis. ALB = serum albumin, APTT = activated partial thromboplastin time, AST = aspartate transaminase, BMI = body mass index, BUN = blood urea nitrogen, CRP = C-reactive protein, ESR = erythrocyte sedimentation rate, FIB = fibrinogen, FOBT = fecal occult blood test, GGT = gamma-glutamyl transpeptidase, PLT = platelet count, PT = prothrombin time, RBC = red blood cell count, RF = random forest, TT = thrombin time, WBC = white blood cell count. The duration of Crohn's disease was measured in months.

of surgery have been identified. Unfortunately, experience has shown that clinicians are often poor judges of disease activity.[13] In addition to the clinician's personal experience, an efficient tool is urgently needed to improve our ability to predict CD-related surgery by integrating subjective and objective evidence of disease activity. As a sub-field of artificial intelligence, ML has the capability of model generalization and high predictive accuracy, and is especially suitable for the analysis of complex data.[21] The increasing use of EMR provides an opportunity for ML to be widely used in diabetic,[22] sepsis,[23] anaphylaxis,[24] colorectal

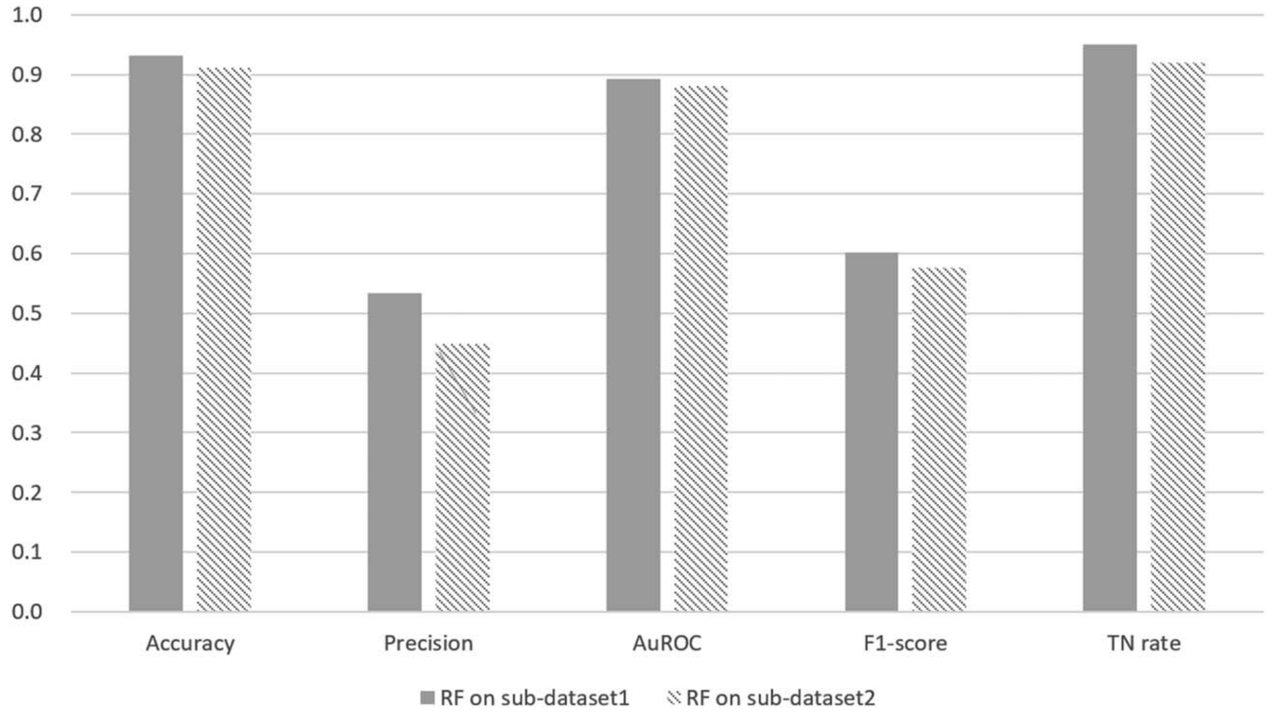

Figure 4. Comparison between RF and LR predictive models performed on sub-dataset 1. RF predictive model performed better than LR model on sub-dataset 1. All metrics represent the average over the 10-folds of the cross validation. AuROC=area under the receiver operating characteristic curve, LR=logic regression, RF=random forest, TN rate=true negative rate.

cancer,[25] lower gastrointestinal bleeding,[26] and other diseases[27,28] as a decision support tool for health-care providers.

In terms of medical treatment for IBD, Waljee et al[29] have already successfully applied a ML algorithm to identify IBD patients on thiopurines with algorithm-predicted objective remission, which has been incorporated into daily clinical use at the University of Michigan. Siegel et al[30] have built another composite algorithm that combined clinical, serological, and genetic variables to provide a better and more accurate prognosis for CD patients. However, their model is only appropriate for patients in earlier stages of the disease and without complications, and genetic testing is not routinely measured, so the poor generalization capability has hindered its use in clinical practice.

There are few models developed to predict CD-related surgery. In recent years, Guizzetti et al[10] used the LR model to predict CD-related surgery and obtained satisfactory results. In our study, we used the same method and similar 10 predictor attributes for evaluation among Chinese CD patients to determine if it was suitable for Chinese patients. At the same time, RF was used for comparison. Compared with LR, RF performed better in prediction on the sub-dataset with these 10 attributes in terms of accuracy, precision, F1 score, TN rate, and the AuROC, which implies that RF predictive model is more suitable for the complex real-world data about Chinese CD patients. In a real-world clinical practice, clinicians would incorporate more clinical information into decision-making. Therefore, RF predictive model was applied on the original dataset again to explore whether the classification accuracy of the

Table 4

Predictive performance of machine learning algorithms on "sub-dataset 2".

Algorithm	Accuracy	Precision	TN rate	F1 score
RF	96.26%	72.13%	97.37%	0.7706
LR	92.33%	49.66%	92.76%	0.6308
DT	95.05%	64.05%	96.24%	0.7112
SVM	92.36%	50.21%	93.00%	0.6288
ANN	90.89%	46.83%	92.24%	0.5757

DT=decision tree, LR=logic regression, RF=random forest, SVM=support vector machine.

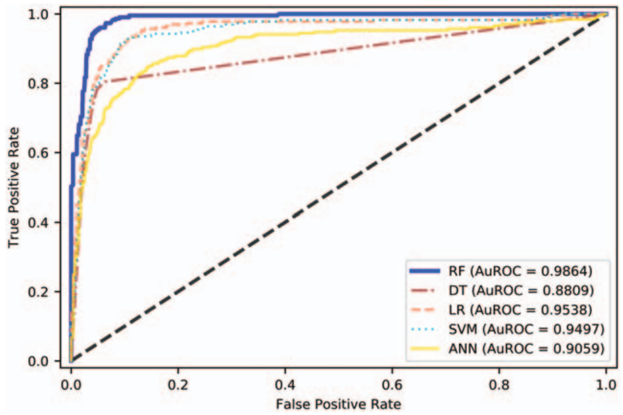

Figure 5. Area under the receiver operating characteristic curve of models on sub-dataset 2. AuROC demonstrates the trade-off between false positive and true positive rates. The closer the curve follows the upper-left corner, the higher the overall accuracy of the model. ANN=artificial neural networks, DT=decision tree, LR=logic regression, RF=random forest, SVM=support vector machine.

ML model could be further improved by reselecting salient predictor attributes. Meanwhile, some of the attributes that have been identified by Guizzetti's LR model[10] were also incorporated into our models and modified to better fit our data. 30 single or aggregated attributes, including *demographic variables*, *clinical performance*, *laboratory tests*, *endoscopic,* and *imaging factors* were reselected by RF algorithms.

In the study by Guizzetti et al,[10] 10 attributes were identified to influence the need for CD-related surgery: *age*, *gender*, *disease location*, *HBI*, *stool frequency*, *immunosuppressive use*, *5-aminosalicylate use*, *presence of a fistula*, *presence of an abscess*, and *presence of an abdominal mass*. Because *HBI* (as a simpler version of *CDAI*) was not available in our original dataset, *CDAI* was used as a substitute for it. The CDAI tool is used to quantify the symptoms of patients with CD and has been confirmed to have a limited role for predicting surgery.[31] The relevance of *CDAI* as a risk factor for surgery, which was evident in Guizzetti et al's LR model,[10] was excluded from sub-dataset 2 by RF-based attribute reselection. But, interestingly, *changes of radiology* were weighted most in the 30 reselected attributes. In data preprocessing, *changes of radiology* from different types of imaging (including MRI and CT enterography, pelvic MRI) were integrated into 5 radiological features, such as normal, inflammatory, fibrotic strictures, fistulas, and abscesses. Consistent with the research results of Jauregui-Amezaga et al,[6] imaging factors could be independent predictors for CD-related surgery in patients.

In our research, we compared the RF predictive model with other predictive models, including LR, SVM, DT, and ANN. The RF model yielded an accuracy of 96.26%, a precision of 72.13%, a TN rate of 97.37%, an F1 score of 0.7706, and the AuROC of 0.9864, which were the highest among all the models under comparison. The classification performance of the 5 different ML algorithms varied with different sub-datasets. Overall, the RF predictive model combined with SMOTE and attribute reselection strategy outperforms the other models under comparison. With its outstanding performance, the result of our model is worthy to be discussed with patients and to be taken into account in therapeutic decision-making in the near future.

5. Conclusion

Unlike previous studies, we used different modeling methods combined with demographic variables, clinical performance, laboratory tests, and endoscopic and imaging factors from real-world data, which made our model more reliable in the routine clinical management of CD. From best of our knowledge, it is the first time for the Chinese patients' data to be used to train a novel RF model to predict patients' risk of experiencing CD-related surgery. However, the proposed models were only developed based on the data acquired from 1 hospital. In future studies, to take advantage of big data for more comprehensive analysis and more accurate prediction, we will deploy this predictive model in the context of a clinical practice to tailor therapeutic strategies to more patients from more hospitals.

Author contributions

Data curation: Qiongxiang Ge, Yuan Dong, Li Xu, Yihong Fan, Ping Xiang, Xuning Gao.

Formal analysis: Qiongxiang Ge, Wenyu Zhang.

Funding acquisition: Yihong Fan.

Methodology: Qiongxiang Ge.

Project administration: Qiongxiang Ge.

Software: Qiongxiang Ge, Yong Chen.

Supervision: Wenyu Zhang.

Writing – original draft: Qiongxiang Ge.

Writing – review & editing: Qiongxiang Ge, Yuan Dong, Wenyu Zhang.

References

[1] Torres J, Mehandru S, Colombel J-F, et al. Crohn's disease. Lancet 2017;389:1741–55.

[2] Bednarz W, Czopnik P, Wojtczak B, et al. Analysis of results of surgical treatment in Crohn's disease. Hepato-gastroenterology 2008;55:998–1001.

[3] Beaugerie L, Seksik P, Larmurier IN, et al. Predictors of Crohn's disease. Gastroenterology 2006;130:650–6.

[4] Solberg IC, Vatn MH, Hoie O, et al. Clinical course in Crohn's disease: results of a Norwegian population-based ten-year follow-up study. Clin Gastroenterol Hepatol 2007;5:1430–8.

[5] Bemelman WA, Warusavitarne J, Sampietro GM, et al. ECCO-ESCP consensus on surgery for Crohn's disease. J Crohns Colitis 2018;12:1–6.

[6] Jauregui-Amezaga A, Rimola J, Ordas I, et al. Value of endoscopy and MRI for predicting intestinal surgery in patients with Crohn's disease in the era of biologics. Gut 2015;64:1397–402.

[7] Singh S, Al-Darmaki A, Frolkis AD, et al. Postoperative mortality among patients with inflammatory bowel diseases: a systematic review and meta-analysis of population-based studies. Gastroenterology 2015;149:928–37.

[8] Silverstein M, Loftus E, Sandborn W, et al. Clinical course and costs of care for Crohn's disease: Markov model analysis of a population-based cohort. Gastroenterology 1999;117:49–57.

[9] Tang J, Rangayyan RM, Xu J, et al. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans Inf Technol Biomed 2009;13:236–51.

[10] Guizzetti L, Zou G, Khanna R, et al. Development of clinical prediction models for surgery and complications in Crohn's disease. J Crohns Colitis 2018;12:167–77.

[11] Cheon JH. Genetics of inflammatory bowel diseases: a comparison between western and eastern perspectives. J Gastroenterol Hepatol 2013;28:220–6.

[12] Zheng JJ, Zhu XS, Huangfu Z, et al. Prevalence and incidence rates of Crohn's disease in mainland China: a meta-analysis of 55 years of research. J Dig Dis 2010;11:161–6.

[13] Gomollón F, Dignass A, Annese V, et al. European evidence-based consensus on the diagnosis and management of Crohn's disease 2016: part 1: diagnosis and medical management. J Crohns Colitis 2017;11:3–25.

[14] Park S, Jeen Y. Role of smoking as a risk factor in east Asian patients with Crohn's disease. Gut Liver 2017;11:7–8.

[15] Ma J, Zhu J, Li N, et al. Severe and differential underestimation of self-reported smoking prevalence in Chinese adolescents. Int J Behav Med 2014;21:662–6.

[16] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[17] Olson DL, Delen D. Advanced Data Mining Techniques. 1st ed. Berlin Heidelberg: Springer –Verlag; 2008:138p.

[18] Sasaki Y. The truth of the F-measure. Teach Tutor Mater 2007;1:1–5.

[19] Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018;16:150.

[20] Dorn SD. Predictors of Crohn's disease. Gastroenterology 2006;131:334–5.

[21] Provost F, Fawcett T. Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. 1st edCA, Sebastopol: O'Reilly Media; 2013. 384p.

[22] Zheng T, Xie W, Xu L, et al. A Machine Learning-based Framework to Identify Type 2 Diabetes Through Electronic Health Records. Int J Med Inform 2017;97:120–7.

[23] Nemati S, Holder A, Razmi F, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018;46:547–53.

[24] Segura-Bedmar I, Colon-Ruiz C, Tejedor-Alonso MA, et al. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. J Biomed Inform 2018;87:50–9.

[25] Kop R, Hoogendoorn M, Teije AT, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med 2016;76:30–8.

[26] Sengupta N, Tapper EB. Derivation and internal validation of a clinical prediction tool for 30-day mortality in lower gastrointestinal bleeding. Am J Med 2017;130:1–8.

[27] Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125:1264–72.

[28] DeGregory KW, Kuiper P, DeSilvio T, et al. A review of machine learning in obesity. Obes Rev 2018;19:668–85.

[29] Waljee AK, Sauder K, Patel A, et al. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. J Crohns Colitis 2017;11:801–10.

[30] Siegel CA, Horton H, Siegel LS, et al. A validated web-based tool to display individualised Crohn's disease predicted outcomes based on clinical, serologic and genetic variables. Aliment Pharmacol Ther 2016;43:262–71.

[31] Rispo A, Imperatore N, Testa A, et al. Combined endoscopic/sonographic-based risk matrix model for predicting one-year risk of surgery: a prospective observational study of a tertiary centre severe/refractory Crohn's disease Cohort. J Crohns Colitis 2018;12:784–93.

[32] Colombel J-F, Panaccione R, Bossuyt P, et al. Effect of tight control management on Crohn's disease (CALM): a multicentre, randomised, controlled phase 3 trial. Lancet 2017;390:2779–89.