




Article

Machine Learning Can Predict the Probability of Biologic Therapy in Patients with Inflammatory Bowel Disease

David Schöler ¹, Karel Kostev ², Maximilian Peters ³, Cosmin Zamfir ⁴, Agnieszka Wolk ⁴, Christoph Roderburg ^{1,†} and Sven H. Loosen ^{1,*,†}

- ¹ Clinic for Gastroenterology, Hepatology and Infectious Diseases, University Hospital Düsseldorf, Medical Faculty of Heinrich Heine University Düsseldorf, Moorenstrasse 5, 40225 Düsseldorf, Germany
- ² Epidemiology, IQVIA, 60549 Frankfurt, Germany
- ³ Real World Insights, IQVIA, 60549 Frankfurt, Germany
- ⁴ Data Science & Advanced Analytics, IQVIA, 60549 Frankfurt, Germany
- * Correspondence: sven.loosen@med.uni-duesseldorf.de; Tel.: +49-211-81-16330
- † These authors share senior authorship.

Abstract: Background: Inflammatory bowel disease (IBD) is of high medical and socioeconomic relevance. Moderate and severe disease courses often require treatment with biologics. The aim of this study was to evaluate machine learning (ML)-based methods for the prediction of biologic therapy in IBD patients using a large prescription database. Methods: The present retrospective cohort study utilized a longitudinal prescription database (LRx). Patients with at least one prescription for an intestinal anti-inflammatory agent from a gastroenterologist between January 2015 and July 2021 were included. Patients who had received an initial biologic therapy prescription (infliximab, adalimumab, golimumab, vedolizumab, or ustekinumab) were categorized as the “biologic group”. The potential predictors included in the machine learning-based models were age, sex, and the 100 most frequently prescribed drugs within 12 months prior to the index date. Six machine learning-based methods were used for the prediction of biologic therapy. Results: A total of 122,089 patients were included in this study. Of these, 15,824 (13.0%) received at least one prescription for a biologic drug. The Light Gradient Boosting Machine had the best performance (accuracy = 74%) and was able to correctly identify 78.5% of the biologics patients and 72.6% of the non-biologics patients in the testing dataset. The most important variable was prednisolone, followed by lower age, mesalazine, budesonide, and ferric iron. Conclusions: In summary, this study reveals the advantages of ML-based models in predicting biologic therapy in IBD patients based on pre-treatment and demographic variables. There is a need for further studies in this regard that take into account individual patient characteristics, i.e., genetics and gut microbiota, to adequately address the challenges of finding optimal treatment strategies for patients with IBD.

Keywords: inflammatory bowel disease; biologics; machine learning; Light Gradient Boosting Machine



Citation: Schöler, D.; Kostev, K.; Peters, M.; Zamfir, C.; Wolk, A.; Roderburg, C.; Loosen, S.H. Machine Learning Can Predict the Probability of Biologic Therapy in Patients with Inflammatory Bowel Disease. *J. Clin. Med.* **2022**, *11*, 4586. <https://doi.org/10.3390/jcm11154586>

Academic Editor: Lorenzo Bertani

Received: 14 June 2022

Accepted: 3 August 2022

Published: 5 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Inflammatory bowel disease (IBD) is of high medical and socioeconomic relevance. There are around 6–8 million cases of IBD worldwide, with prevalence rising in industrialized countries [1]. Moderate and severe disease courses often require the use of immunosuppressive drugs, including azathioprine and TNF-alpha-, JAK1/3-, and IL12/23-inhibitors, to maintain remission. A variety of risk factors for the development of IBD have been identified in the past [2], such as genetic factors, e.g., mutations in *NOD2* for Crohn's disease, or changes in the gut microbiota [3]. However, the pathogenesis of IBD is still largely unknown, and there is a lack of predictive criteria for a clinical response to a particular therapy. In clinical routine, many years of therapy with alternating therapy regimens are often needed to find the appropriate substances for individual patients. Depending on the

severity of the disease, this is also associated with an increased risk of cancer, especially in individuals with ulcerative colitis, and the need for surgery, which represents a significant morbidity/mortality risk and also significantly reduces patient quality of life.

Recently, machine learning (ML) algorithms have been increasingly used to support clinical decision making, e.g., in predicting the mortality of septic patients via extreme gradient boosting (XGBoost; [4]), or in the assessment of myocardial ischemia using angiography with a Light Gradient Boosting Machine (GBM) [5]. The aim of this study was to evaluate ML-based methods for the prediction of biologic therapy in patients with Crohn's disease and ulcerative colitis using a large prescription database.

2. Materials and Methods

2.1. Data Sources

The present retrospective cohort study was based on the IQVIA LRx database [6]. This database comprises about 80% of prescriptions reimbursed by statutory health insurance funds in Germany. Data are available at the patient level and include information on patient age and sex. In accordance with data privacy legislation, all patient information is fully anonymized by the data provider. Each prescription is available with the full product information (e.g., brand, substance, package size, and product form) and dates dispensed. The database does not contain diagnoses or details of laboratory tests. The LRx database has been used in various previous studies [7]. As the LRx database contains prescription information, but no diagnosis information, data from the Disease Analyzer database (IQVIA) were used for the present study to define Crohn's disease (CD) and ulcerative colitis (UC) cases based on the prescribed therapy. The Disease Analyzer database covers around 3% of primary-care practices in Germany. This database includes demographic, diagnosis, and prescription data obtained in an anonymized format from computer systems used in general and specialized practices. Previous research has shown that this is representative of primary-care practices in Germany [8].

2.2. Study Population and Outcome

Using data from the Disease Analyzer database, we first compared prescribed therapies in patients with CD/UC attending gastroenterologist practices with those in patients without CD/UC. The most significant difference was observed for intestinal anti-inflammatory agents (ATC: A07E excluding vedolizumab), with prescription rates of 80% in CD/UC patients and 2.0% in non-CD/UC patients. Finally, we included all patients from the LRx database with at least one prescription for an intestinal anti-inflammatory agent from a gastroenterologist between January 2015 and July 2021. Patients who received an initial biologic therapy prescription (infliximab, adalimumab, golimumab, vedolizumab, or ustekinumab) were categorized as the "biologic group". These patients were classified as early adopters (patients whose first biologics prescription was issued at least 180 days prior to the last record in the database) and late adopters (patients whose first biologics prescription was within the time period of 0–180 days before the last record in the database). The date of the first prescription for biologics was considered the index date. Patients who did not receive any biologics during the study period were classified as the "non-biologic group". For these patients, the last prescription for any drug documented in the database was considered the index date. One further criterion for inclusion was an observation time of at least 18 months (prior to the last database record).

2.3. Potential Predictors and Statistical Analyses

The objective of the ML model was to differentiate between patients who received biologics and those who did not, and to identify key predictors of biologic therapy initiation. Potential predictors included in the machine learning-based models corresponded to prescription data obtained prior to the index date. More than 400 different drugs prescribed within 12 months prior to the index date were found. The model took into account the

100 most frequently prescribed drugs as well as patient age and sex, and prescriptions were normalized to the total number of prescriptions per patient.

Six machine learning-based methods for the prediction of biologic therapy were tested: Light Gradient Boosting Machine (LGBM) [9], linear support vector classification (SVC) [10], logistic regression [11], cat boost [12], competitive gradient descent (CGD) [13], and random forest [14].

Training of the ML systems was completed using data from non-biologics patients and early adopters. The dataset was split into two sets: 80% for training and 20% for testing. The LRx database used provides a snapshot of each patient's prescription history. It was therefore possible that a patient had not yet received biologics but would receive them in the future. We tested whether a subset of the misclassified non-biologics patients, i.e., false-positive patients, might be future biologics patients. This experiment was performed by excluding late adopters from the training set and testing them separately. To compensate for the imbalanced dataset, the class weight for the biologics patients was set to the ratio between non-biologics and biologics patients.

3. Results

3.1. Characteristics of the Study Sample and Incidence of Biologic Therapy

A total of 122,089 patients were included in this study. The mean (standard deviation) age of the study sample was 51.3 (SD: 18.5) years, and 55.7% of patients were female. The five most frequently prescribed therapies within 12 months prior to the index date were mesalazine (55.9%), budesonide (35.2%), pantoprazole (27.5%), metamizole (27.0%), and prednisolone (24.5%). As intestinal anti-inflammatory agents were used for the definition of the study population, they were not considered a co-therapy. Of the 122,089 IBD patients, 15,824 (13.0%) received at least one prescription for a biologic drug. Biologics patients were younger (43.9 (SD: 15.8) years) than non-biologics patients (52.9 (SD: 17.5) years). The proportion of women was slightly lower among the biologics patients (53.6% vs. 56.5%).

3.2. Performance of the Biologic Therapy Prediction Models

Figure 1 outlines the performance of the biologic therapy prediction models. The performance of LGBM (accuracy = 74%) was highest, followed by LSVC (72%) and logistic regression (72%).

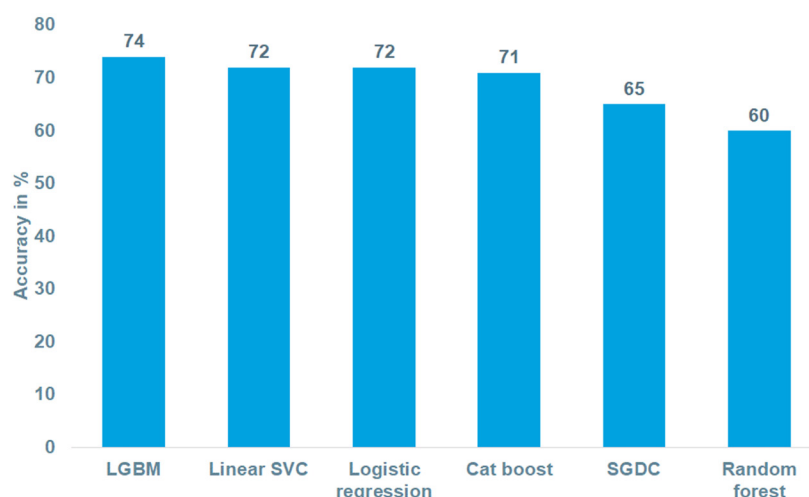


Figure 1. Accuracy of algorithms evaluated.

Abbreviations: LGBM—Light Gradient Boosting Machine; SVC—Support Vector Classifier; SGDC—Stochastic Gradient Descent Classifier. The F1 score was calculated using the training dataset.

In terms of sensitivity, the LGBM model was able to correctly identify 78.5% of the biologics patients and 72.6% of the non-biologics patients in the testing dataset (Figure 2).

The majority of the patients who were classified as biologic in the training dataset but who had not received biologics at this point were correctly classified as biologics patients in the test dataset containing late adopters. The chosen algorithm was therefore able to identify patients for whom traditional therapies were failing and for whom biologics would be used in the future.

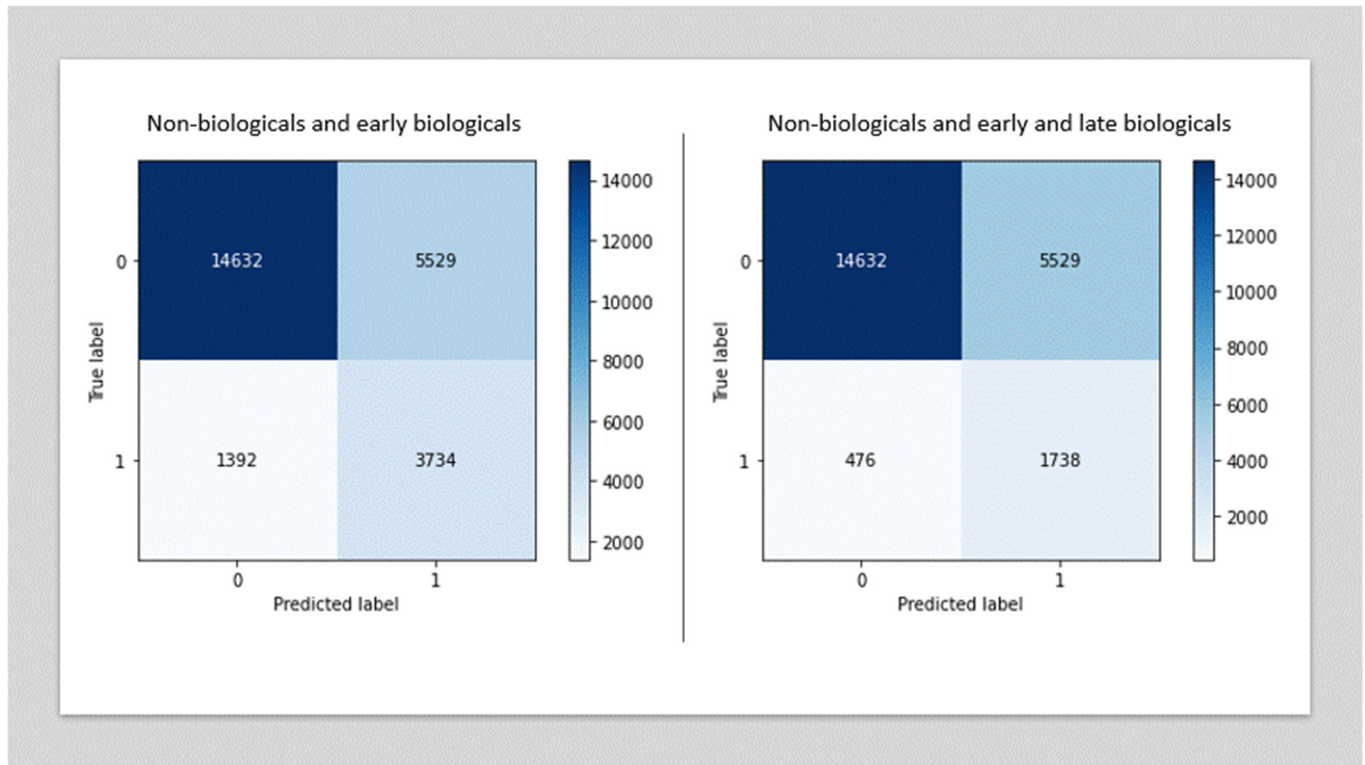


Figure 2. Predicted and true biologics users (LGBM classifier).

The prediction of biologics patients was slightly less accurate when late adopters were included (72.9%) (Figure 2).

3.3. Most Important Variables Predicting the Probability of Biologic Therapy

Table 1 displays the therapies with the largest differences in proportions between patients with and without biologic therapy. Calcium products (12.3% vs. 3.9%), immunosuppressants (mainly azathioprine, 33.3% vs. 11.0%), iron products (24.7% vs. 8.6%), and oral corticosteroids (71.1% vs. 26.1%), as well as antidiarrheal microorganisms (4.7% vs. 2.0%), were the therapy classes with the biggest differences.

Figure 3 shows the feature importance of the variables from the LGBM model. The most significant variable was prednisolone, followed by lower age, mesalazine, budesonide, and ferric iron. These variables were followed by the use of azathioprine, metamizole sodium, colecalciferol, cefuroxime axetil, and pantoprazole. Of note, the use of prednisolone, budesonide, and azathioprine is part of first-line therapy in the treatment of IBD. A much higher value was found for prednisolone when compared to the other variables in the top 10.

Table 1. Therapies with largest proportional differences between patients with and without biologic therapy.

Therapy Class	Proportion among Biologics Patients in % (n = 15,284)	Proportion among Non-Biologics Patients in % (n = 106,265)	Prevalence Ratio (Biologics/Non-Biologics)	p-Value
Calcium	12.3	3.9	3.2	<0.001
Other immunosuppressants *	33.3	11.0	3.0	<0.001
Iron	24.7	8.6	2.9	<0.001
Oral corticosteroids	71.1	26.1	2.7	<0.001
Antidiarrheal microorganisms	4.7	2.0	2.4	<0.001
Intestinal corticosteroids	64.2	33.4	1.9	<0.001
Vitamin D	20.1	10.9	1.8	<0.001
Oral fluoroquinolones	10.4	6.9	1.5	<0.001
Gastroprokinetics	6.9	4.7	1.5	<0.001
Antitussives	6.8	4.7	1.4	<0.001
PPI	54.0	41.4	1.3	<0.001

* Mainly azathioprine.

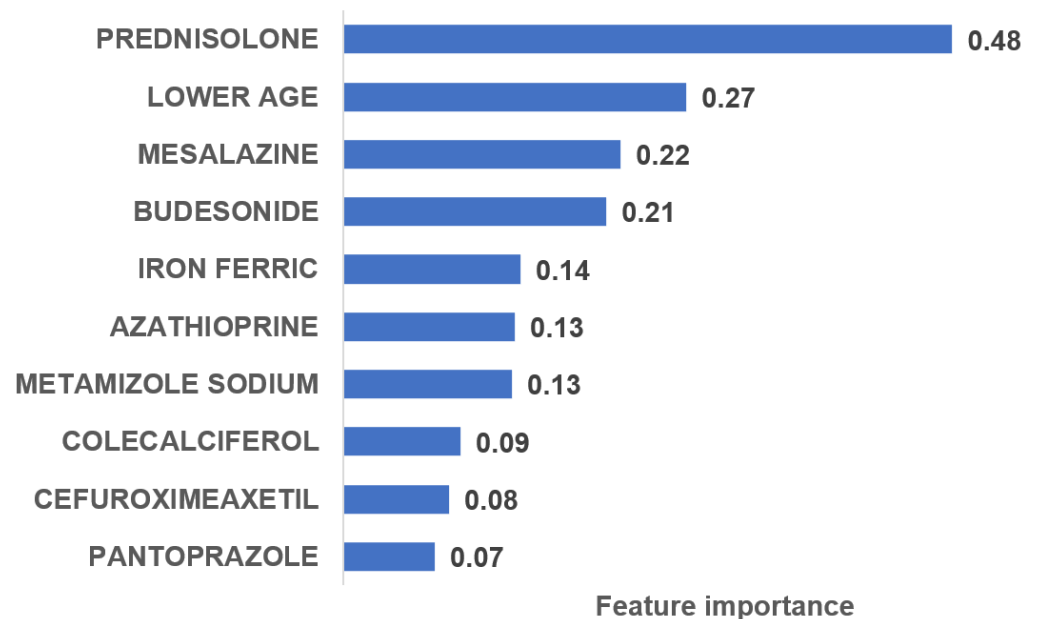


Figure 3. Feature importance in the LGBM model (top 10 variables).

4. Discussion

In the present study, six ML-based models were tested to determine their accuracy in the prediction of biologic therapies in a large cohort of more than 120,000 IBD patients. All the ML models tested had a high level of accuracy, with LGBM achieving the highest F1 values. The prescriptions of corticosteroids, mesalazine, and ferric iron, as well as lower age, were the variables with the highest feature importance.

Our data suggest that biologics tend to be considered more often for the treatment of younger patients. Some studies have demonstrated that clinical features and the extent of disease in IBD patients appear to be more severe with younger age [15,16]. As biologics are usually prescribed for patients with severe disease, this may explain the role of younger age as one of the most significant predictors. In their large study completed in South Korea, Choi et al. reported that the odds of receiving biologic drugs were 2.3 times higher in the early-onset group than in the late-onset group [17].

A high proportion of patients received oral corticosteroids, intestinal corticosteroids, and other immunosuppressants (mainly azathioprine), but also iron prior to biologic therapy, and these therapies were important predictor features in the ML model. The prescription of these drugs is common practice. For example, vedolizumab and ustekinumab have been approved for the treatment of patients with moderate to severe UC or CD who exhibit an inadequate response, loss of response, or intolerance to corticosteroids and immunosuppressives [18].

Iron is usually given to patients with anemia, which is a common extraintestinal manifestation of IBD. Patients with IBD commonly have iron-deficiency anemia due to chronic blood loss and impaired iron absorption as a result of tissue inflammation [19]. This complication also demonstrates the severity of IBD, which can impact the patient's chances of receiving a biologic drug.

Vitamin D (colecalfiferol) was one of the top 10 most commonly prescribed medications in the ML model. Vitamin D deficiency is common in patients with active IBD [20] and is linked to disease activity, more frequent relapses, and higher postoperative recurrence [21]. A further symptom of severe IBD is pain, which is a long-standing problem for the majority of the patients affected [22]. Metamizole, which was also one of the top ten medications in the ML model, is an analgesic often prescribed in IBD patients.

Although cefuroxime axetil also ranked among the top ten most common medications in the model, it was found to have lower future significance in predicting the prescription of biologics. This antibiotic drug is sometimes used to treat bacterial infections that may arise due to complications of IBD and which may cause fistulas [23].

This study is subject to a number of limitations. First, patients were selected based on the prescription of intestinal anti-inflammatory drugs, which presents the possibility of selection bias. Although the administration of these drugs reflects the common guidelines for the treatment of patients with CD/UC, a number of patients do not receive them, e.g., because of intolerance. Second, the diagnoses of CD and UC were based on database queries [6,8] and were not standardized by the same clinical criteria, i.e., by the use of a positive histology. Moreover, in the LRx database, no information about diagnoses was available and no stratification into CD and UC was possible. Third, the database did not include data on therapy duration and daily dose for the majority of patients. Fourth, the precision of the ML models was not tested in different subgroups while taking into account the severity of the disease. Fifth, the external test sets were not available. Sixth, the retrospective analysis only shows associations and does not show any causal relationships.

However, as other working groups suggest [24,25], our study shows that in combination with big data, ML approaches can help improve the treatment of IBD patients in the future.

In summary, this investigation studied the feasibility of ML-based models in predicting future biologic therapy in IBD patients based on pre-treatment and demographic variables. Further studies are required that take into account individual patient characteristics, i.e., genetics and gut microbiota [2,3], to adequately address the challenges of finding optimal treatment strategies for patients with IBD.

Author Contributions: M.P., A.W. and C.Z. designed the study, performed the statistical analyses, and generated figures and tables. D.S., K.K., C.R. and S.H.L. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: No specific funding was received for the completion of this study. In general, work in the lab of T.L. was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program through the ERC Consolidator Grant PhaseControl (Grant Agreement n° 771083). The lab of T.L. was further supported by German Cancer Aid (Deutsche Krebshilfe 110043 and a Mildred Scheel Professorship) and the German Research Foundation (SFB-TRR57/P06, LU 1360/3-1, CRC1380/A01, and CA 830/3-1).

Informed Consent Statement: The "Disease Analyzer" and LRx database, used for analysis in this study, contain anonymized electronic patient records. Patient data were analyzed in aggregated form

without individual data becoming available. Individual consent forms were not obtained in line with national and European legislation.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest: K.K., M.P., C.Z., und A.W. are employees of IQVIA. The authors declare that they have no competing interests.

References

- Alatab, S.; Sepanlou, S.G.; Ikuta, K.; Vahedi, H.; Bisignano, C.; Safiri, S.; Sadeghi, A.; Nixon, M.R.; Abdoli, A.; Abolhassani, H.; et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **2020**, *5*, 17–30. [[CrossRef](#)]
- Ananthakrishnan, A.N. Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* **2015**, *12*, 205–217. [[CrossRef](#)] [[PubMed](#)]
- Turpin, W.; Goethel, A.; Bedrani, L.; Croitoru, K. Determinants of IBD Heritability: Genes, Bugs, and More. *Inflamm. Bowel Dis.* **2018**, *24*, 1133–1148. [[CrossRef](#)] [[PubMed](#)]
- Hou, N.; Li, M.; He, L.; Xie, B.; Wang, L.; Zhang, R.; Yu, Y.; Sun, X.; Pan, Z.; Wang, K. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. *J. Transl. Med.* **2020**, *18*, 462. [[CrossRef](#)]
- Hae, H.; Kang, S.-J.; Kim, W.-J.; Choi, S.-Y.; Lee, J.-G.; Bae, Y.; Cho, H.; Yang, D.H.; Kang, J.-W.; Lim, T.-H.; et al. Machine learning assessment of myocardial ischemia using angiography: Development and retrospective validation. *PLoS Med.* **2018**, *15*, e1002693. [[CrossRef](#)]
- Richter, H.; Dombrowski, S.; Hamer, H.; Hadji, P.; Kostev, K. Use of a German longitudinal prescription database (LRx) in pharmacoepidemiology. *Ger. Med. Sci.* **2015**, *13*, Doc14. [[CrossRef](#)]
- Helwig, U.; Braegger, F.; Kostev, K.; Schmidt, C. Comparative Analysis of the 3-Year Persistence Rate with Second-Line Vedolizumab and Tumor Necrosis Factor- α Inhibitors in Patients with Inflammatory Bowel Disease Followed in Gastroenterology Practices in Germany. *Dig. Dis.* **2020**, *38*, 466–473. [[CrossRef](#)]
- Rathmann, W.; Bongaerts, B.; Carius, H.-J.; Kruppert, S.; Kostev, K. Basic characteristics and representativeness of the German Disease Analyzer database. *Int. J. Clin. Pharmacol. Ther.* **2018**, *56*, 459–466. [[CrossRef](#)]
- Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inform. Process. Syst.* **2017**, *30*, 3146–3154.
- Sidey-Gibbons, J.A.M.; Sidey-Gibbons, C.J. Machine learning in medicine: A practical introduction. *BMC Med. Res. Methodol.* **2019**, *19*, 64. [[CrossRef](#)]
- Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **2002**, *35*, 352–359. [[CrossRef](#)]
- Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for big data: An interdisciplinary review. *J. Big Data* **2020**, *7*, 94. [[CrossRef](#)] [[PubMed](#)]
- Sum, J.; Leung, C.S.; Ho, K. A Limitation of Gradient Descent Learning. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *31*, 2227–2232. [[CrossRef](#)]
- Rigatti, S.J. Random Forest. *J. Insur. Med.* **2017**, *47*, 31–39. [[CrossRef](#)]
- Cosnes, J.; Bourrier, A.; Nion-Larmurier, I.; Sokol, H.; Beaugerie, L.; Seksik, P. Factors affecting outcomes in Crohn's disease over 15 years. *Gut* **2012**, *61*, 1140–1145. [[CrossRef](#)]
- Lee, J.H.; Cheon, J.H.; Moon, C.M.; Park, J.J.; Hong, S.P.; Kim, T.I.; Kim, W.H. Do patients with ulcerative colitis diagnosed at a young age have more severe disease activity than patients diagnosed when older? *Digestion* **2010**, *81*, 237–243. [[CrossRef](#)]
- Choi, Y.I.; Kim, Y.J.; Chung, J.W.; Kim, K.O.; Kim, H.; Park, R.W.; Park, D.K. Effect of age on the initiation of biologic agent therapy in patients with inflammatory bowel disease: Korean common data model cohort study. *JMIR Med. Inform.* **2020**, *8*, e15124. [[CrossRef](#)]
- Sulz, M.C.; Burri, E.; Michetti, P.; Rogler, G.; Peyrin-Biroulet, L.; Seibold, F. Treatment Algorithms for Crohn's Disease. *Digestion* **2020**, *101*, 43–57. [[CrossRef](#)]
- Kaitha, S.; Bashir, M.; Ali, T. Iron deficiency anemia in inflammatory bowel disease. *World J. Gastrointest. Pathophysiol.* **2015**, *6*, 62–72. [[CrossRef](#)]
- Ham, M.; Longhi, M.S.; Lahiff, C.; Cheifetz, A.; Robson, S.; Moss, A.C. Vitamin D levels in adults with Crohn's disease are responsive to disease activity and treatment. *Inflamm. Bowel Dis.* **2014**, *20*, 856–860. [[CrossRef](#)]
- Nielsen, O.H.; Hansen, T.I.; Gubatan, J.M.; Jensen, K.B.; Rejnmark, L. Managing Vitamin D deficiency in inflammatory bowel disease. *Frontline Gastroenterol.* **2019**, *10*, 394–400. [[CrossRef](#)] [[PubMed](#)]
- Zeit, J.; Ak, M.; Muller-Mottet, S.; Scharl, S.; Biedermann, L.; Fournier, N.; Frei, P.; Pittet, V.; Scharl, M.; Fried, M.; et al. Pain in IBD patients: Very frequent and frequently insufficiently taken into account. *PLoS ONE* **2016**, *11*, e0156666. [[CrossRef](#)] [[PubMed](#)]

23. Lal, S.; Steinhart, A.H. Antibiotic therapy for Crohn's disease: A review. *Can. J. Gastroenterol.* **2006**, *20*, 651–655. [[CrossRef](#)] [[PubMed](#)]
24. Choi, Y.I.; Park, S.J.; Chung, J.W.; Kim, K.O.; Cho, J.H.; Kim, Y.J.; Lee, K.Y.; Kim, K.G.; Park, D.K.; Kim, Y.J. Development of machine learning model to predict the 5-year risk of starting biologic agents in patients with inflammatory bowel disease (Ibd): K-cdm network study. *J. Clin. Med.* **2020**, *9*, 3427. [[CrossRef](#)] [[PubMed](#)]
25. Olivera, P.; Danese, S.; Jay, N.; Natoli, G.; Peyrin-Biroulet, L. Big data in IBD: A look into the future. *Nat. Rev. Gastroenterol. Hepatol.* **2019**, *16*, 312–321. [[CrossRef](#)] [[PubMed](#)]