Annals of Gastroenterology (2026) **39,** 247-253

Machine learning in predicting treatment response and remission in inflammatory bowel disease: a systematic review

Sheza Malik[a], Renisha Redij[b], Dushyant Singh Dahiya[c], Chengu Niu[d], Douglas G. Adler[e]

Emory University Hospital, Atlanta, Georgia; Trinity Health Livonia Hospital, Michigan; University of Kansas, Kansas City; University of Nebraska, Omaha, Nebraska, USA; Advent Health, Denver, Colorado, USA

Abstract

Background The heterogeneity of inflammatory bowel disease (IBD) and its unpredictable course have always been a challenge for gastroenterologists, with regard to predicting the disease response using endoscopic techniques. Machine learning (ML) models have shown some early promise in predicting treatment response in IBD patients.

Methods We conducted a systematic review of studies investigating the application of ML to predict treatment response and remission in IBD patients. We used the CHARMS checklist for data extraction. Bias was assessed with the PROBAST tool.

Results We included in our review 6 studies that evaluated numbers of IBD patients ranging from 67 to 3004. ML models demonstrated low to moderate predictive accuracy for treatment response and remission (area under the receiver operating characteristic curve: 0.489-0.811; sensitivity: 0.46-0.96; specificity: 0.56-0.98). The studies that utilized ML models with more input variables performed better. Furthermore, only 2 studies performed external validation, and half of the studies demonstrated a substantial risk of bias due to missing data/overfitting, and variability in outcome definition

Conclusions ML models show considerable promise in predicting treatment outcomes and remission in IBD. However, given the substantial bias in studies so far, future studies should use a standardized methodology, external validation, and an interpretable broader input variable.

Keywords Machine learning, inflammatory bowel disease, treatment, monitoring, response

Ann Gastroenterol 2026; 39 (2): 247-253

[a]Gastroenterology and Hepatology, Emory University Hospital, Atlanta, Georgia, USA (Sheza Malik, Chengu Niu); [b]Internal Medicine, Trinity Health Livonia Hospital, Michigan, USA (Renisha Redij); [c]Gastroenterology and Hepatology, University of Kansas, Kansas City, Kansas, USA (Dushyant Singh Dahiya); [d]Gastroenterology and Hepatology, University of Nebraska, Omaha, Nebraska, USA (Chengu Niu) [e]Gastroneterology and Hepatology, Center for Advanced Therapeutic Endoscopy, Centura Health, Denver, Colorado, USA (Douglas G. Adler)

Conflict of Interest: DGA: Consultant, Boston Scientific. All other authors: No conflict of interest

Correspondence to: Douglas G. Adler, MD, FACG, AGAF, FASGE, Director, Center for Advanced Therapeutic Endoscopy, Advent Health, Denver, CO, USA, e-mail: dougraham2001@gmail.com

Received 13 April 2025; accepted 24 December 2025; published online 12 February 2026

DOI: https://doi.org/10.20524/aog.2026.1041

Introduction

Inflammatory bowel disease (IBD), comprising Crohn's disease (CD), ulcerative colitis (UC), and unclassified IBD, has a heterogeneous nature, making it very challenging for gastroenterologists to predict treatment remission and response on the basis of endoscopic scores [1,2]. The unpredictable response to various commonly used pharmacological therapies further complicates clinical management [3]. Machine learning (ML) models have shown early promise to overcome these challenges.

ML models utilize large and diverse datasets to recognize meaningful patterns in clinical, biochemical, and imaging data [4]. The methods include supervised, unsupervised, and ensemble learning algorithms. ML models have shown potential in predicting treatment response and remission in IBD patients [5]. These models have utilized multiple input variables (electronic health records, serum markers, genomics, radiographic, histological, and endoscopic data) to predict the disease activity and response to the treatment.

Our study aimed to evaluate the ML models' performance in predicting treatment response and remission in patients with

IBD. By examining the methodological quality and predictive accuracy of the ML models, we aimed to assess the current performance of ML models for IBD.

Materials and methods

In our systematic review, we followed the Preferred Reporting Items for Systematic Reviews and Meta-analysis [PRISMA] statement [6]. The details of the PRISMA checklist are provided in Supplementary Table 1.

Literature search

We conducted a comprehensive database search in February 2025, including Ovid Medline, Ovid EMBASE, Scopus, and Web of Science. With input from our team, an experienced medical librarian designed the search strategy, using controlled vocabulary and specific keywords for ML, artificial intelligence, IBD, CD, UC, treatment response prediction, clinical remission, and biologic therapy.

Eligibility criteria

Studies were selected based on predefined inclusion and exclusion criteria. No *a priori* exclusions were applied based on the type of predictors. Studies incorporating clinical, biochemical, endoscopic, imaging, or genetic variables were eligible, provided that these features were available at or before the treatment decision. Only peer-reviewed journal articles published in English were considered. In addition, in our review, we considered any predictive model utilizing data-driven optimization, regularization, or non-linear pattern recognition beyond standard statistical regression (e.g., penalized regression methods) as ML models.

Exclusion criteria included studies that used traditional statistical models and studies that focused only on IBD diagnosis. Additionally, case reports, conference abstracts, and review articles without primary data were excluded.

Selection process

Two authors, SM and RR, independently reviewed the titles and abstracts of studies returned by the primary search. Studies that did not address the research question were excluded. The full texts of the remaining articles were then examined. Any discrepancies in selecting articles were resolved through consensus and discussion with another co-author, DSD.

Data extraction and quality assessment

In accordance with the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS checklist) [7], 2 authors, SM and RR, independently extracted data. Baseline characteristics were gathered, including year of publication, study design, patient population, and sample size. Data on ML methods, model validation strategies, performance metrics, and input variables—including clinical markers, biochemical parameters, endoscopic scores, genetic features, and imaging findings—were also gathered.

Risk of bias assessment

The study co-authors (SM, RR) assessed each study for risk of bias according to the TRIPOD recommendations (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) [8], using the PROBAST tool (Prediction model Risk Of Bias Assessment Tool) [9]. Studies were assessed across 4 key areas: (i) participants; (ii) predictors; (iii) outcome; and (iv) analysis. For each domain, the risk of bias and applicability to the intended clinical setting were evaluated. Discrepancies were resolved through consensus among the coauthors.

Statistical analysis

Because of the heterogeneity in ML models, input variables, definition of the outcome and type/strategy for validation, a meta-analysis could not be conducted. Instead, a descriptive synthesis was performed. We also separated the studies based on validation (internal vs. external validation), allowing for an assessment of model generalizability [10].

Results

Study selection

A total of 29 studies were identified by our search strategy, of which 6 were included in the final analysis [9-14]. There was a high degree of agreement between the 2 reviewers (SM, RR) regarding the inclusion of studies (Cohen's k: 0.977, 95% confidence interval: 0.81-1.00). Fig. 1 shows the PRISMA flowchart for the study identification and selection.

Characteristics of studies

Our review encompasses 6 retrospective studies that investigated ML in predicting the Treatment Response and Remission in IBD [11-16]. These studies include 3 retrospective analyses and 3 that reported a *post hoc* ML analysis of Phase III multicenter, randomized placebo-controlled clinical trials. Participant numbers ranged from 67 to 3004. The average age was 31 years. The baseline characteristics of these studies are summarized in Table 1.

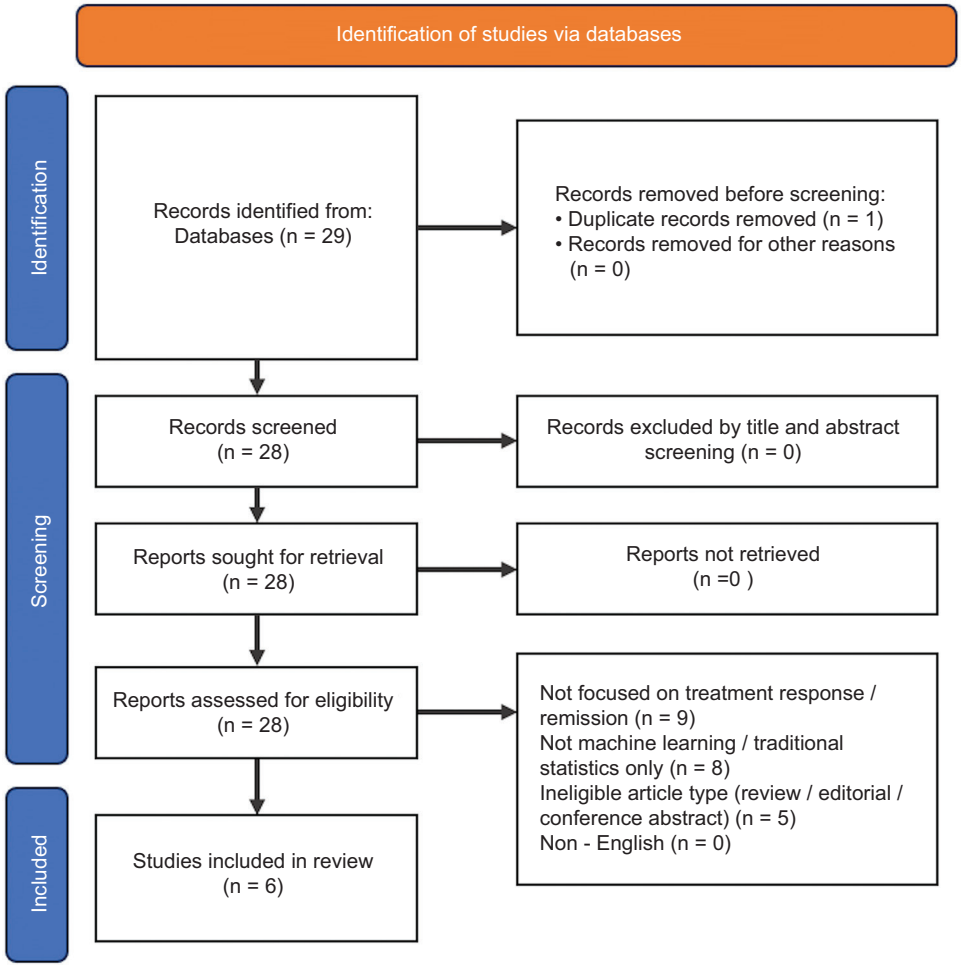

Figure 1 PRISMA flowchart
From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, *et al* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71
For more information, visit: http://www.prisma-statement.org/

Table 1 Study characteristics and validation details

Study [ref.]	Year	Drug studied	ML algorithm	Validation type	Cohort details/Sample size
Chen *et al* [11]	2022	Vedolizumab	Elastic Net	Internal	543 patients total — VISIBLE 1 trial: 160 patients; VARSITY trial: 383 patients
Harun *et al* [12]	2024	Etrolizumab	XGBoost	Internal	3004 patients screened; 1684 enrolled in induction phases; 463 received treatment during maintenance
Miyoshi *et al* [13]	2021	Vedolizumab	Random Forest	External	Training cohort: 34 patients from Kyorin University Hospital. Validation cohort: 35 patients from Toho University Sakura Medical Center
Morikubo *et al* [14]	2024	Ustekinumab	Logistic Regression	External	Training cohort: 25 patients from Kyorin University Hospital, Tokyo. Test cohort: 46 patients from Toho University Sakura Medical Center, Chiba
Qiu *et al* [15]	2024	Infliximab	XGBoost	Internal & External	746 patients from 3 centers — FAH (379), SAH (176), SYF (191)
Waljee *et al* [16]	2018	Vedolizumab	Random Forest	Internal	472 subjects from a phase III clinical trial on vedolizumab; original cohort n=594, with exclusions for missing data or no active inflammation

ML, machine learning; XGBoost, extreme gradient boosting; FAH, The First Affiliated Hospital, Sun Yat-sen University; SAH, The Sixth Affiliated Hospital, Sun Yat-sen University; SYF, Sir Run Run Shaw Hospital, Zhejiang University

ML approaches

The included studies employed a range of supervised learning algorithms, with a predominance of ensemble methods, particularly RF and XGBoost. None of the included studies used deep learning models, probably due to the limited availability of large-scale datasets.

Input features used in ML models

The included studies varied in their selection of predictive input features. Commonly used biochemical markers included C-reactive protein, fecal calprotectin and white blood cell count. Clinical variables such as age, disease duration, prior treatment failure and medication history were frequently incorporated into machine-learning models. Some studies included endoscopic findings, using scoring systems such as the Simple Endoscopic Score for Crohn's Disease, Ulcerative Colitis Endoscopic Index of Severity scores, and the Mayo Endoscopic Score.

Performance of ML models

In our review, ML models showed variable performance across studies (area under the receiver operating characteristic curve [AUROC] 0.489-0.811), with AUROC 0.70-0.79 interpreted as acceptable/moderate and ≥0.80 as strong/high [17].

The ensemble learning algorithms (e.g., Random Forest and XGBoost) consistently performed better than simpler models, probably because of their ability to recognize complex, nonlinear relationships. In our review, Sensitivity ranged from 0.462-0.964 and specificity from 0.56-0.98. The details of various ML models are summarized in Table 2.

Internal validation was performed in 4 studies, whereas only 2 studies implemented external validation. The handling of missing data was inconsistent and variable across studies. The patient population (CD vs. UC) and definitions of treatment response/remission also varied across included studies, while different therapeutic agents were assessed.

Risk of bias and methodological limitations

In the Participants domain of the PROBAST tool, all studies (6/6: 100%) were considered as having a low risk of bias. The concerns regarding applicability, however, were unclear in 4/6 studies (66%) (insufficient information on inclusion and exclusion criteria).

In the Predictors domain of the PROBAST tool, all studies clearly defined predictor variables. However, 2 studies (33%) [11,12] had an unclear risk of bias, as they lacked sufficient information on whether predictors were assessed independently of outcome data. The concern regarding the applicability of the predictor domain was low across all 6 studies.

The Outcomes domain showed a low risk of bias and low applicability concerns in all 6 studies (100%). For the Analysis domain, substantial concerns were identified. Three of 6 studies (50%) showed a high risk of bias, due to unclear handling of continuous and categorical variables, and a lack of information on how missing data, class imbalance, and potential non-linearity/time dependence were managed. None of the studies addressed potential overfitting in their models. The PROBAST details are provided in Table 3.

Discussion

Our study highlights the potential role of ML for predicting treatment response and remission in people with IBD. In 6 studies, ML models showed good prediction performance, especially when multivariate clinical and biochemical input variables, such as biomarkers, endoscopic scores, and genetic markers, were used. However, the current clinical applicability of the methodology and the potential for its extensive adoption are still limited, given the large heterogeneity.

Previous systematic reviews in IBD have examined the potential for ML as a tool to aid in the diagnosis of IBD, with limited investigation into ML's role in predicting treatment response and remission in patients already diagnosed with IBD [18,19]. We assessed ML models designed to predict therapy response and remission in IBD. Of all the algorithms evaluated, the ensemble ML models, Extreme Gradient Boosting (XGBoost) and Random Forest, surpassed simpler linear models, further substantiating their suitability in prognostic modeling in oncology as well as other chronic disorders [20]. Nevertheless, none of the studies used explainable artificial intelligence (XAI) methods. XAI methods show how complex models make predictions. A popular scheme, SHapley Additive exPlanations (SHAP), shows how much each variable contributes to each predictor, making model outputs clearer. In other areas, such as cardiovascular risk estimation, XAI has greater clinician confidence [21], highlighting its significance for complex conditions like IBD.

In our review, approximately half of the included studies had a high risk of bias, mainly due to inadequate statistical methodology. This finding is consistent with previous studies, which have highlighted the importance of transparent model reporting as outlined in the TRIPOD statement [22]. Most studies in our review were retrospective, single-center studies, raising concerns about overfitting and generalizability [23]. Additionally, heterogeneity in the definitions of clinical response and remission (such as clinical vs. biochemical), and in outcome measures, further complicates cross-study comparisons [24].

Future research should prioritize 3 key directions. First, prospective multicenter trials following the TRIPOD framework [8] are essential for validating ML models in real-

Table 2 Performance of ML models

Study [ref.]	Outcome (s) to be predicted/Def of outcome	Sensitivity	Specificity	AUROC
Chen *et al* [11]	Treatment outcomes of vedolizumab in ulcerative colitis Def: clinical outcome, clinical remission at Week 52 (defined as a complete Mayo score of ≤2 points and no individual subscore >1 point); and HRQoL outcome, IBDQ remission at Week 52 (defined as IBDQ total score >170).	For test set Clinical remission ENRR: 0.474 RF: 0.474 IBDQ remission ENRR: 0.964 RF: 0.589	For test set Clinical remission ENRR: 0.868 RF: 0.838 IBDQ remission ENRR: 0.160 RF: 0.560	Test set Clinical remission ENRR: 0.811 RF: 0.726 IBDQ remission ENRR: 0.721 RF: 0.593
Harun *et al* [12]	First, to identify the key prognostic factors impacting remission, and second, to perform a complete exposure-response (E-R) analysis using a comprehensive list of explanatory variables and/or potential confounders in the E-R relationship.	Induction phase model: 0.82 (0.72-0.92) Maintenance phase model: 0.85 (0.71-0.99)	Induction phase model: 0.68 (0.56-0.80) Maintenance phase model: 0.72 (0.56-0.88)	Induction phase model: 0.74 (0.68-0.80) Maintenance phase model: 0.75 (0.63-0.87)
Myoshi *et al* [13]	Efficacy of vedolizumab to achieve steroid-free clinical remission at week 22 in ulcerative colitis	0.87 (0.72-1)	0.67 (0.51-0.83)	0.76 (0.6-0.92)
Morikubo *et al* [14]	Efficacy of ustekinumab (UST) to achieve steroid-free clinical remission at week 22 in ulcerative colitis (UC) Def: Steroid-free clinical remission (SFCR) at week 22 was evaluated as the clinical outcome. SFCR was defined as a Lichtiger index of 4 or lower. Patients who terminated UST treatment or needed surgery because of insufficient control of UC inflammation before week 22 were regarded as not achieving clinical remission.	-	-	-
Qiu *et al* [15]	Primary outcome: prediction of long-term clinical remission at 52 weeks and beyond 2 years in Crohn's disease after starting infliximab. The secondary outcome was predicting short-term clinical remission at 26 weeks.		26 week For SAH: XGB: 0.988 LR: 0.986 Lasso/EN: 0.988 For SYF: XGB: 0.845 LR: 0.810 Lasso/EN: 0.845	26 week For SAH: XGB: 0.733 (0.47-0.97) LR: 0.519 (0.31-1.00) Lasso/EN: 0.497 (0.351-0.643) For SYF: XGB: 0.517 (0.42-0.65) LR: 0.584 (0.46-0.65) Lasso/EN: 0.598 (0.50-0.70)
Waljee *et al* [16]	Prediction of corticosteroid-free biologic remission at week 52 in Crohn's disease patients treated with vedolizumab Def The primary outcome was corticosteroid-free biologic remission with vedolizumab at week 52, defined by no use of steroid medications (including prednisone and budeso nide) at week 52, and reduction of CRP from >5 mg/L at baseline to ≤5 mg/L at week 52.	Baseline model: 0.64 Week 6 model: 0.76 Simplified model: 0.73	Baseline model: 0.56 Week 6 model: 0.71 Simplified model: 0.69	Baseline model: 0.65 (0.53-0.77) Week 6 model: 0.75 (0.64-0.86) Simplified model: 0.75 (0.70-0.81)

ML, machine learning; AUROC, area under the receiver operating characteristic curve; ENRR, elastic net regularized regression; RF, random forest; HRQoL, health-related quality of life; IBDQ, Inflammatory Bowel Disease Questionnaire; SAH, The Sixth Affiliated Hospital, Sun Yat-sen University; LR, logistic regression; EN, elastic net; SYF, Sir Run Run Shaw Hospital, Zhejiang University; XGB, extreme gradient boosting; CRP, C-reactive protein

world IBD populations. Second, reaching consensus on core biomarkers and remission criteria will harmonize predictors and endpoints across studies, facilitating external validation and meta-analytic synthesis [25]. Third, integrating explainability tools should enhance clinicians' understanding and build confidence in ML-generated outputs, rather than substituting for methodological rigor [26]. Despite moderate accuracy, existing models are largely limited to predicting response to individual biologic agents (e.g., vedolizumab or infliximab). The next generation of models must move beyond single-drug predictions to enable comparative forecasting across multiple therapeutic options, thereby supporting precision therapy selection in routine clinical practice.

In conclusion, ML represents a potentially transformative opportunity to advance precision medicine in IBD. Its clinical integration will depend on rigorous prospective validation, adoption of standardized reporting frameworks, and close interdisciplinary collaboration among data scientists, clinicians, and methodologists to bridge the gap between algorithmic innovation and tangible patient benefit.

Table 3 PROBAST scoring

Author [ref.]	Risk of Bias				Applicability			Overall	
	1. Participants	2. Predictors	3. Outcome	4. Analysis	1. Participants	2. Predictors	3. Outcome	Risk of Bias	Applicability
Chen *et al* [11]	Low	Unclear	Low	High	Unclear	Low	Low	High	Low
Harun *et al* [12]	Low	Unclear	Low	High	Unclear	Low	Low	High	Low
Myoshi *et al* [13]	Low	Low	Low	Unclear	Unclear	Low	Low	Unclear	Low
Morikubo *et al* [14]	Low	Low	Low	Unclear	Unclear	Low	Low	Unclear	Low
Qiu *et al* [15]	Low	Low	Low	High	Low	Low	Low	High	Low
Waljee *et al* [16]	Low	Low	Low	Low	Low	Low	Low	Low/ Unclear	Low

PROBAST domains (Participants, Predictors, Outcome, Analysis)

Summary Box

What is already known:

- Predicting treatment response in inflammatory bowel disease (IBD) is difficult, in view of the variability of the disease
- Traditional tools like endoscopy and biomarkers have limited predictive power
- Machine learning (ML) has shown promise in IBD diagnosis, but its role in predicting remission is less explored
- Existing studies lack consistency in methods and validation

What the new findings are:

- ML models, especially Random Forest and XGBoost, show moderate-to-good accuracy in predicting IBD treatment outcomes
- Multivariate data sets improve model performance
- Only a third of studies used external validation; half had a high risk of bias
- Standardization and prospective validation are key for clinical use

References

1. Vasudevan A, Gibson PR, van Langenberg DR. Time to clinical response and remission for therapeutics in inflammatory bowel diseases: what should the clinician expect, what should patients be told? *World J Gastroenterol* 2017;**23**:6385-6402.
2. D'Incà R, Sturniolo G. Biomarkers in IBD: what to utilize for the diagnosis? *Diagnostics (Basel)* 2023;**13**:2931.
3. Plevris N, Lees CW. Disease monitoring in inflammatory bowel disease: evolving principles and possibilities. *Gastroenterology* 2022;**162**:1456-1475.
4. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering (Basel)* 2023;**10**:1435.
5. Kraszewski S, Szczurek W, Szymczak J, Reguła M, Neubauer K. Machine learning prediction model for inflammatory bowel disease based on laboratory markers. Working model in a discovery cohort study. *J Clin Med* 2021;**10**:4745.
6. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* 2021;**10**:89.
7. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;**11**:e1001744.
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;**350**:g7594.
9. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;**170**:51-58.
10. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;**22**:276-282.
11. Chen J, Girard M, Wang S, Kisfalvi K, Lirio R. Using supervised machine learning approach to predict treatment outcomes of vedolizumab in ulcerative colitis patients. *J Biopharm Stat* 2022;**32**:330-345.
12. Harun R, Lu J, Kassir N, Zhang W. Machine learning-based quantification of patient factors impacting remission in patients with ulcerative colitis: insights from etrolizumab phase III clinical trials. *Clin Pharmacol Ther* 2024;**115**:815-824.
13. Miyoshi J, Maeda T, Matsuoka K, et al. Machine learning using clinical data at baseline predicts the efficacy of vedolizumab at week 22 in patients with ulcerative colitis. *Sci Rep* 2021;**11**:16440.
14. Morikubo H, Tojima R, Maeda T, et al. Machine learning using clinical data at baseline predicts the medium-term efficacy of ustekinumab in patients with ulcerative colitis. *Sci Rep* 2024;**14**:4386.
15. Qiu Y, Hu S, Chao K, et al. Developing a machine-learning prediction model for infliximab response in Crohn's disease: integrating clinical characteristics and longitudinal laboratory trends. *Inflamm Bowel Dis* 2025;**31**:1334-1343.
16. Waljee AK, Wallace BI, Cohen-Mekelburg S, et al. Development and validation of machine learning models in prediction of remission in patients with moderate to severe Crohn disease. *JAMA Netw Open* 2019;**2**:e193721.
17. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Wiley; 2013.
18. Zulqarnain F, Rhoads SF, Syed S. Machine and deep learning in inflammatory bowel disease. *Curr Opin Gastroenterol* 2023;**39**:294-300.

19. Pei J, Wang G, Li Y, et al. Utility of four machine learning approaches for identifying ulcerative colitis and Crohn's disease. *Heliyon* 2024;**10**:e23439.
20. Stidham RW, Takenaka K. Artificial intelligence for disease assessment in inflammatory bowel disease: how will it change our practice? *Gastroenterology* 2022;**162**:1493-1506.
21. Al-Droubi SS, Jahangir E, Kochendorfer KM, et al. Artificial intelligence modelling to assess the risk of cardiovascular disease in oncology patients. *Eur Heart J Digit Health* 2023;**4**:302-315.
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;**350**:g7594.
23. Aliferis C, Simon G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. 2024 Mar 5. In: Simon GJ, Aliferis C, editors. Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls [Internet]. Cham (CH): Springer; 2024.
24. Ma C, Panaccione R, Fedorak RN, et al. Heterogeneity in definitions of endpoints for clinical trials of ulcerative colitis: a systematic review for development of a core outcome set. *Clin Gastroenterol Hepatol* 2018;**16**:637-647.
25. Bova G, Domenichiello A, Letzen JE, et al. Developing consensus on core outcome sets of domains for acute, the transition from acute to chronic, recurrent/episodic, and chronic pain: results of the INTEGRATE-pain Delphi process. *EClinicalMedicine* 2023;**66**:102340.
26. Agrawal R, Gupta T, Gupta S, Chauhan S, Patel P, Hamdare S. Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagn Pathol* 2025;**20**:105.

Supplementary material

Supplementary Table 1 PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
		TITLE	
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	3
		INTRODUCTION	
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
		METHODS	
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	5-7
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5-7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5-7
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5-7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5-7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	5-7
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I2) for each meta-analysis.	5-7
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6-7
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6-7
		RESULTS	
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7-9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7-9
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-9
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	7-9

(*Contd...*)

Supplementary Table 1 (*Continued*)

Section/topic	#	Checklist item	Reported on page #
Synthesis of results	21	Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency	7-9
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	7-9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	7-9
		DISCUSSION	
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	10-12
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	10-12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	10-12
		FUNDING	
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1-2

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6 (7): e1000097. doi: 10.1371/journal.pmed1000097