Published prediction models cannot replace ileocolonoscopy for monitoring mucosal disease activity in Crohn's disease patients

- a systematic review and external validation of published prediction models

Short title

Endoscopic activity prediction in Crohn's disease

Authors

Eelco C. Brand[1,2], MD, Sjoerd G. Elias[3], MD, PhD, Itta M. Minderhoud[4], MD, PhD, Julius J. van der Veen[1], BSc, LLB, Filip J. Baert[5], MD, PhD, David Laharie[6],MD, PhD, Peter Bossuyt[7], MD, Yoram Bouhnik[8], MD, PhD, Anthony Buisson[9], MD, Guy Lambrecht[10], MD, Edouard Louis[11], MD, PhD, Benjamin Pariente[12], MD, PhD, Marieke J. Pierik[13], MD, PhD, C. Janneke van der Woude[14], MD, PhD, Geert R.A.M. D'Haens[15], MD, PhD, Séverine Vermeire[16], MD, PhD, Bas Oldenburg[1], MD, PhD. On behalf of the Dutch Initiative on Crohn and Colitis (ICC).

Affiliations

1. Department of Gastroenterology and Hepatology, University Medical Center Utrecht, Utrecht, The Netherlands.
2. Laboratory for Translational Immunology, University Medical Center Utrecht, Utrecht, The Netherlands.
3. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
4. Department of Gastroenterology and Hepatology, Tergooi hospitals, Blaricum/Hilversum, The Netherlands.
5. Department of Gastroenterology, AZ Delta, Roeselare, Belgium
6. Service d'Hépato-gastroentérologie et Oncologie Digestive, Hôpital Haut-Lévêque, Bordeaux, France
7. IBD Clinic, Imelda General Hospital, Bonheiden, Belgium
8. Department of Gastroenterology, Beaujon Hospital, APHP, Paris Diderot University, Clichy, France
9. Department of Gastroenterology, Estaing University Hospital, Clermont-Ferrand, France
10. Department of Gastroenterology, AZ Damiaan, Oostende, Belgium
11. Department of Gastroenterology, Liège University Hospital CHU, Liège, Belgium
12. Department of Gastroenterology, Huriez Hospital, Lille 2 University, Lille, France
13. Department of Gastroenterology and Hepatology, Maastricht University Medical Center, Maastricht, The Netherlands
14. Department of Gastroenterology and Hepatology, Erasmus Medical Center, Rotterdam, The Netherlands

15. Department of Gastroenterology, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
16. Department of Gastroenterology and Hepatology, University Hospitals Leuven, Leuven, Belgium

Grant support

Eelco Brand is supported by the Alexandre Suerman program for MD and PhD candidates of the University Medical Center Utrecht, Netherlands.

Abbreviations

AUC, area under the receiver operating characteristic curve
CD, Crohn’s disease
CDAI, Crohn’s disease activity index
CDEIS, Crohn’s disease endoscopic index of severity
CHARMS, critical appraisal and data extraction for systematic reviews of prediction modeling studies
CI, confidence interval
CRP, C-reactive protein
EH, endoscopic healing
ESR, erythrocyte sedimentation rate
HBI, Harvey-Bradshaw Index
SESCD, Simple endoscopic score for Crohn’s disease
TAILORIX, a randomized controlled trial investigating tailored treatment with infliximab for active luminal Crohn’s disease
TNF-α, tumor necrosis factor-α
UAI, Utrecht Activity Index

Corresponding author

Bas Oldenburg, MD, PhD, Department of Gastroenterology and Hepatology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands. E-mail: b.oldenburg@umcutrecht.nl.

Disclosures

ECB is supported by the Alexandre Suerman program for MD and PhD candidates of the University Medical Center Utrecht, Netherlands, and is co-applicant on a received Investigator Initiated Research Grants research grant of Pfizer.
IMM served on an advisory panel of Ferring.
FB received research grants from Abbvie, Chiesi, Ipsen, Mundipharma, Roche and Takeda and Speakers and consultancy fees from Abbvie, Cellgene, Ferring, Janssen, Mundipharma, MSD, Pfizer, Takeda, Vifor.
DL received board and lecture fees from Abbvie, Biogaran, Biogen, Ferring, HAC-pharma, Janssen, MSD, Novartis, Pfizer, Prometheus, Roche, Takeda, Theradiag, Tillots.
PB has received educational grants from AbbVie, Mundipharma, Pfizer, Janssen; speaker fees from AbbVie, Takeda, Pfizer; and advisory board fees from Hospira, Janssen, MSD, Mundipharma, Roche, Pfizer, Sandoz, Takeda, Abbvie.

AB declares consulting fees for Abbvie, Hospira, Takeda and lecture fees from Abbvie, Hospira, Takeda, MSD, Vifor Pharma, Sanofi-Aventis, Ferring.
EL has received the following fees: Research Grant: Takeda, Pfizer, Janssen; Educational Grant: Abbvie, MSD, Takeda, Janssen; Speaker Fees: Abbvie, Ferring, MSD, Falk, Takeda, Hospira, Janssen, Pfizer; Advisory Board: Abbvie, Ferring, MSD, Takeda, Celgene, Hospira, Janssen, Pfizer ; Consultant: Abbvie.
BP received board and lecture fees from Abbvie, Biogaran, FerringJanssen, MSD, Pfizer, Takeda, and Lilly.
CJvW received grants from Abbvie, Jansen, Pfizer, Tramedico, Falk. Speakers and consultancy fees from Abbvie, Falk, Ferring, Janssen, Mundipharma, MSD, Pfizer, Takeda, Vifor.
GRAMD has served as advisor for Abbvie, Ablynx, Allergan, Amakem, Amgen, AM Pharma, Arena Pharmaceuticals, AstraZeneca, Avaxia, Biogen, Bristol Meiers Squibb, Boerhinger Ingelheim, Celgene/Receptos, Celltrion, Cosmo, Covidien/Medtronics, Ferring, DrFALK Pharma, Eli Lilly, Engene, Galapagos, Genentech/Roche, Gilead, Glaxo Smith Kline, Hospira/Pfizer, Immunic, Johnson and Johnson, Lycera, Medimetrics, Millenium/Takeda, Mitsubishi Pharma, Merck Sharp Dome, Mundipharma, Nextbiotics, Novonordisk, Otsuka, Pfizer/Hospira, Photopill, Prometheus laboratories/Nestle, Progenity, Protagonist, Robarts Clinical Trials, Salix, Samsung Bioepis, Sandoz, Seres/Nestle, Setpoint, Shire, Teva, Tigenix, Tillotts, Topivert, Versant and Vifor; received speaker fees from Abbvie, Biogen, Ferring, Johnson and Johnson, Merck Sharp Dome, Mundipharma, Norgine, Pfizer, Samsung Bioepis, Shire, Millenium/Takeda, Tillotts and Vifor.
SV is senior clinical researcher of the Research Foundation Flanders (FWO). SV received grant support from MSD, Abbvie, Pfizer, J&J and Takeda, lecture fees from Abbvie, MSD, Ferring Pharmaceuticals, Takeda, Hospira, Tillotts, J&J and consultancy fees from Abbvie, Takeda, Pfizer, Ferring Pharmaceuticals, Shire Pharmaceuticals Group, MSD,Tillotts, Prodigest, Ablynx, Robarts Clinical Trials, Prometheus, Progenity, Hospira, Mundipharma, Celgene, Galapagos, Genentech/Roche.
BO reports grants from MSD, Abbvie, Takeda, Cablon, Ferring, Falk, and Pfizer.
SGE, JJvdV, YB, GL, MJP declare no conflicts of interest.

Writing assistance
None.

Author contributions
ECB, SGE, BO: conception and design of the study.
ECB, JJVdV, SGE, BO: performance of systematic review.
IMM, FJB, DL, PB, YB, AB, GL, EL, BP, MJP, CJvdW, GRAMD, SV, BO: generation and acquisition of data.
ECB, SGE, BO: analysis and interpretation of the data.
ECB, SGE, BO: drafting of the manuscript.
All authors: critical revision of the manuscript for important intellectual content.
All authors: approval of the final version of the manuscript.

Acknowledgments

We would like to thank the following authors for providing additional data: Y. Bosi, M. de Bruyn, E. Domènech, P. Eder, M.T. Herranz Bachiller, M.-A. Meuwis, P. Miranda-García, S. Nancey, G. Opdenakker, J. Panés, J.C. Preiss, E. Stragier, A. Viscido. None received compensation outside of their usual salary.

Word count

Abstract: 260 words

Manuscript: 5901 words

ABSTRACT

Background & aims Endoscopic healing (EH), an important therapeutic target in Crohn's Disease (CD), requires ileocolonoscopy, which is costly and burdensome. We aimed to determine whether published non-invasive prediction models could replace ileocolonoscopy for monitoring CD activity.

Methods We performed a systematic review of all published diagnostic models predicting endoscopic activity or EH in CD. We externally validated these models for the outcome endoscopic activity (Crohn's disease endoscopic index of severity≥3) in the TAILORIX (346 ileocolonoscopies in 155 patients) and the Utrecht Activity Index (UAI) (93 ileocolonoscopies in 82 patients) dataset. As benchmark, we assessed the performance of fecal calprotectin (FC) and C-reactive protein (CRP) as single biomarkers.

Results After screening 5303 titles, 27 models (21 studies) were identified. Seven models could be externally validated, for which the area under the receiver operating characteristic curves (AUCs) [95%-confidence interval] ranged from 0.61 [0.51-0.70] to 0.81 [0.76-0.86] (TAILORIX) and from 0.58 [0.39-0.76] to 0.82 [0.73-0.91] (UAI). The AUCs for FC were 0.79 [0.74-0.85] and 0.82 [0.73-0.92], and for CRP 0.72 [0.66-0.77] and 0.80 [0.71-0.88]. A threshold yielding a positive predictive value ≥90% could be identified for 4/7 models, FC and CRP, and yielding a negative predictive value (NPV) ≥90% for 2/7 models but not for FC and CRP. Most ileocolonoscopies (TAILORIX:66.5%, UAI:72.6%) could correctly be avoided using FC ≤100 and >250μg/g, however, at the cost of many incorrectly avoided ileocolonoscopies (TAILORIX:18.7%, UAI:19.8%).

Conclusions Published prediction models cannot sufficiently predict endoscopic activity in CD, especially due to low NPVs. Therefore, ileocolonoscopy remains the mainstay for mucosal disease activity assessment in CD.

Keywords

Inflammatory bowel disease; mucosal healing; colonoscopy.

INTRODUCTION

Endoscopic healing (EH), i.e. mucosal healing assessed by endoscopy, has become the new therapeutic goal in the treatment of Crohn's disease.[1–3] EH is associated with long-term corticosteroid-free remission and a decreased risk of surgery and hospitalisations.[4] Ileocolonoscopy plays a key role in monitoring EH, but is time-consuming, costly and potentially burdensome.

A non-invasive prediction model, reliably predicting endoscopic activity, would therefore be of great benefit in clinical practice. Ileocolonoscopies could be avoided if absence or presence of endoscopic activity is predicted with sufficient certainty. Treatment can be concluded to be effective in case of correctly predicted EH and therapeutic changes might be needed for correctly predicted endoscopic activity. Taking into consideration that an avoided ileocolonoscopy for incorrectly predicted EH might lead to undertreatment, while an avoided ileocolonoscopy for incorrectly predicted endoscopic activity might lead to unneeded escalation of therapy.

Symptom-based scores, such as the Crohn's disease activity index (CDAI)[5] and the Harvey-Bradshaw Index (HBI)[6], have been found to correlate poorly with endoscopic findings.[7,8] Routinely used biomarkers in Crohn's disease management, such as C-reactive protein (CRP) and fecal calprotectin are not deemed reliably enough to replace ileocolonoscopy.[7,9] Several non-invasive prediction models, combining multiple predictors, have been developed and published, but almost none are externally validated, which is considered essential before clinical implementation.[10]

We therefore aimed to determine whether published, non-invasive prediction models can replace ileocolonoscopies for assessment of endoscopic disease activity in Crohn's disease. To this end, we performed a systematic review to identify all published prediction models, and subsequently externally validated these models in two different prospective cohorts of Crohn's disease patients.

MATERIALS AND METHODS

Systematic review

Our systematic review adheres to the critical appraisal and data extraction for systematic reviews of prediction modeling studies (CHARMS)-checklist[11] (Supplementary Material 1), and was prospectively registered in the international prospective registry for systematic reviews (PROSPERO): CRD42018092633**.**

Search strategy

A comprehensive systematic search was performed in PubMed, Embase, and the Cochrane Library from inception until February 14, 2018. The search strategy consisted of title/abstract and MeSH or Emtree terms for Crohn's disease and EH or endoscopic activity[12] (Supplementary Material 2). The modified Ingui filter[13,14] was applied to specifically search for prediction models.

Eligibility criteria

We included articles based on the following eligibility criteria: 1) full-text articles published in peer-reviewed journals; 2) no language restrictions; 3) a cross-sectional study design (i.e. the developed model must predict the endoscopic outcome at the same moment in time); 4) studies exclusively aimed at adult Crohn's disease patients, or, if both ulcerative colitis and Crohn's disease patients were studied, enabling the extraction of data for Crohn's disease alone; 5) description of the prediction/diagnostic model, risk score, non-invasive index, clinical decision rule, or equivalent for the outcome ileocolonic endoscopic activity or EH, assessed by (ileo)colonoscopy and quantified by the Crohn's Disease Index of Endoscopic Severity (CDEIS)[15], Simple Endoscopic Scale for Crohn's Disease (SES-CD)[16], Rutgeerts score[17] or another systematic endoscopic activity index; 6) models including at least three predictors. The exclusion criteria are provided in more detail in Figure 1 and Supplementary Material 3.

Study selection

Two authors (ECB and JJvdV) independently screened all titles and abstracts for eligibility and subsequently assessed the full-texts of the remaining articles for final inclusion. Thereafter reference lists of included articles and reviews were crosschecked for additional potentially relevant articles, until no further publications were identified. Agreement was achieved in consensus meetings between the two authors. Any disagreement was resolved by consulting two other authors (SGE and BO).

Data extraction

The following data were extracted independently by two authors (ECB and JJvdV) from included studies: 1) study characteristics, 2) patient characteristics, 3) ileocolonoscopy characteristics, 4) model development, 5) model specifications, 6) model performance in original study (Supplementary Material 4). If the prediction model was not presented in sufficient detail to be validated authors were requested to provide additional data.

Critical appraisal

The risk of bias and applicability of the included studies was independently assessed by ECB and JJvdV based on the Prediction model Risk Of Bias ASsessment Tool (PROBAST).[18] Publications were assessed for risk of bias on four domains (participant selection, predictors, outcome and analyses) and for applicability on three domains (participant selection, predictors, outcome).

Data sources

Two separate datasets were used for the external validation of the included models.

TAILORIX dataset

The multicenter (27 centers in Belgium, France and the Netherlands) randomized TAILORIX trial aimed to explore the role of tailored treatment with infliximab in biological naïve patients with active luminal Crohn's disease.[19] In short, after screening 167 patients, 122 patients with moderate to severely active Crohn's disease (CDAI 220-450), were started on infliximab in combination with an immunomodulator. Patients were randomized to receive one of three regimens of monitoring-based dosage adjustments, based on clinical symptoms only or on a combination of clinical symptoms, CRP-levels, fecal calprotectin levels, and infliximab serum trough levels.

At week 0, 12, and 54, patients underwent a pre-scheduled ileocolonoscopy. The endoscopic activity of disease was scored based on the Crohn's Disease Endoscopic Index of Severity (CDEIS)[15]. The CDEIS was independently scored by physicians blinded to patient and clinical information, based on videos of the ileocolonoscopy when available. For our current study we included all patients with at least one ileocolonoscopy, irrespective of the outcome of the baseline screening and included for each patient the available ileocolonoscopies at all timepoints.

Utrecht activity index dataset

The Utrecht Activity Index (UAI) study aimed to develop and externally validate a model that could predict the CDEIS.[20] Here, we use the development dataset of the UAI study including 82 consecutive Crohn's disease patients undergoing 93 ileocolonoscopies in a Dutch tertiary care center. Patients with an ileo- or colostomy or those with a history of major intestinal surgery were excluded. The CDEIS score was assessed by one endoscopist and, in 72 out of 93 procedures, reviewed by a second endoscopist, blinded for patient details. Since the intraclass correlation was found to be high (0.86)[20], the CDEIS scores from the first endoscopist were used for the present study.

Predictors and outcome in external validation

Matching predictors for validation

We matched the predictors from the identified models to the variables in our validation cohorts. In the TAILORIX dataset we used the predictor values obtained as shortly before or after the ileocolonoscopies as possible for the present analysis. In the UAI dataset all predictors, e.g.

blood and fecal tests, were obtained before ileocolonoscopy and thus blinded for the endoscopy results. Matching of a few noteworthy predictors is discussed in more depth in Supplementary Material 4.

Outcome

The CDEIS was used as the reference standard. This score ranges from 0-44 increasing with the severity of disease activity.[15] A CDEIS≥3 was considered endoscopic activity[3], hence EH was defined as a CDEIS<3.[21,22] We used endoscopic activity as outcome in all analyses, irrespective of the outcomes used to originally develop the different models. If a model was originally developed to predict EH, we inversed the weights of the predictors to predict endoscopic activity.

Statistical analyses

Statistical analyses were performed using R language environment for statistical computing 3.5.1 for Mac.[23] Baseline characteristics of both cohorts are depicted as numbers and proportions for categorical variables, and medians and boundaries of the interquartile range (IQR) for continuous variables.

Missing data

Missing data for predictors, if present, ranged from 0.3% (smoking behavior) to 17.3% (fecal calprotectin) in the TAILORIX and 1.1% (CDAI) to 8.6% (fecal calprotectin) in the UAI dataset. The CDEIS was scored if ileocolonoscopy was performed and thus no outcome data were missing. Missing data were handled by creating multiple imputed datasets (31 for the TAILORIX, and 12 for the UAI dataset, reflecting the proportion of incomplete cases per database[24]) by iterative (25 iterations) chained equations employing the MICE package.[25] Analyses were performed in all multiple imputed datasets and subsequently pooled by Rubin's rules.[26]

Discrimination

To assess the discriminatory abilities of the included models we estimated the area under the receiver operating characteristic curve (AUC) with 95%-confidence intervals (95%-CIs) in both validation datasets. We corrected for clustering of up to 3 ileocolonoscopies per patient employing the Obuchowski method.[27]

Fecal calprotectin and CRP as benchmark

Because fecal calprotectin and CRP are the most commonly used biomarkers in clinical practice for the evaluation of disease activity in Crohn's disease, we decided to benchmark model performance against the performance of these biomarkers as continuous variables. Furthermore, we evaluated the performance of CRP and fecal calprotectin employing commonly used clinical thresholds,[9] i.e.: fecal calprotectin >100 µg/g or >250 µg/g, and CRP >5 (in the TAILORIX dataset) or >7 (lower detection limit of CRP in the UAI dataset) mg/L.

Calibration

We assessed the calibration, i.e. the extent to which the predicted outcome is in line with the observed outcome, for models either yielding a predicted probability for endoscopic activity or models predicting CDEIS continuously. To this end we constructed calibration plots using restricted cubic splines in mixed generalized linear models to account for clustering of ileocolonoscopies within patients. Because no models except one[28] reported the intercept, we assessed calibration after intercept updating (i.e. recalibration-in-the-large).[29] Subsequently, we assessed the calibration following logistic recalibration,[29] and linear calibration for the model predicting CDEIS continuously. Updating the intercept adjusts the prediction towards the prevalence in the new setting, while with logistic (or linear) recalibration a slope correction is additionally calculated for which all weights per predictor are equally corrected. A slope correction <1 and >1 indicate overestimation and underestimation, respectively, of the predicted probability, while 1 reflects a perfect fit for the original model. Calibration could not be assessed for models generating a dichotomous outcome (i.e. no activity or activity).

Scatter plots for CRP and fecal calprotectin versus CDEIS

The correlation between the observed CDEIS and continuous fecal calprotectin and CRP levels were assessed by the construction of scatter plots. The log of the CRP was used to improve the readability of the scatter plot. The curves were based on a cubic spline-based function in mixed generalized linear models to correct for clustering within patients. The explained variance in observed CDEIS by the individual biomarkers was assessed through the marginal R^2-values.

Correct avoidance of ileocolonoscopies

Lastly, we assessed whether, based on the predictions, ileocolonoscopies could safely be avoided. Therefore, we sought to identify thresholds in the predicted probabilities or predicted CDEIS yielding an NPV and PPV ≥90% or, if not attainable, ≥80%. Ileocolonoscopies could safely be avoided for expected EH if a patient scores below the lower threshold (based on the NPV), or for expected endoscopic activity if a value above or equal to the upper threshold (based on the PPV) is predicted. In patients scoring in between these thresholds ileocolonoscopy would still be required. Subsequently, based on the identified thresholds we calculated the proportion including Wilson 95%-CIs[30] of ileocolonoscopies that would have been correctly avoided (i.e. EH predicted and EH observed or endoscopic activity predicted and endoscopic activity observed), incorrectly avoided or that still needed to be performed. Based on the identified thresholds we also calculated the NPV, PPV, sensitivity, specificity, and overall accuracy including Wilson 95%-CIs[30].

We performed these analyses for the included models, fecal calprotectin and CRP levels as single continuous biomarkers, and commonly used fecal calprotectin thresholds (<100 μg/g or >250 μg/g).[9] Because the calibration of all models benefited from logistic or linear recalibration, we used the recalibrated model formulas in the analyses for potentially avoidable ileocolonoscopies.

RESULTS

Systematic literature search

After screening 5,303 publications, we identified 21 eligible studies[20,28,31–49] reporting on 27 diagnostic models for the prediction of ileocolonic endoscopic activity and/or EH in patients with Crohn's disease (Figure 1 and Supplementary Material 3).

Study and model characteristics

The characteristics of identified studies and models are depicted in table 1 and Supplementary Materials 5-9. CRP (17/27 [63.0%]), fecal calprotectin (13/27 [48.1%]) and the HBI (5/27 [18.5%]) were the most frequently used predictors (Table 1 and Supplementary Material 5). The outcome definition for endoscopic activity, when assessed as a dichotomous outcome, varied among studies from a SES-CD of >0 to >7, a CDEIS of >3 and ≥3, and a (modified) Rutgeerts' score of ≥i1 to ≥i2b (Supplementary Material 6). Discrimination and calibration measures were described for 11/27 (40.7%) and 6/27 (22.2%) of models (Supplementary material 9). In only one study the developed model was validated internally as well as externally[20]. In no other studies validation was performed.

Risk of bias assessment

Overall, all developed models were at high risk of bias, mainly because of lack of correction for optimism in the model estimates, a low number of outcomes relative to the number of predictors considered, and because most studies did not fully evaluate the performance of the models (Supplementary Materials 10-11). Because most studies did not report whether predictor and outcome assessment were blinded, most models scored an unclear risk of bias in these domains. If the model was judged to be at high risk of inapplicability this was mostly because the model had been developed for a subgroup of Crohn's disease patients (e.g. after ileocecal resection) or the outcome was not purely ileocolonoscopic but based on ileocolonoscopy and/or another test (e.g. computed tomography).

Models in- and excluded for external validation

Ten studies[28,34,35,37,39,42,44,45,48] reporting on 15 models did not provide sufficient information to apply these models. Authors from these studies were contacted, and extra information was provided by the authors of four studies[28,34,37,45] representing five models. In total, 20 models could not be externally validated, because: 1) ten models[35,39,40,42,44,48] did not provide sufficient detail to allow external validation, 2) six models[31,34,37,45,46] included investigational biomarkers, 3) four models[33,38,41,49] used predictors not available in the TAILORIX and/or UAI dataset (Supplementary material 12). Nonetheless, seven models[20,28,32,36,41,43,47] could be externally validated: six models in the TAILORIX dataset, and five models in the UAI dataset (obviously, the UAI model was not validated in its own development dataset).

Characteristics of models included for external validation

Three models[20,32,43] were developed in Crohn's disease patients in general, the Garcia-Planella and Herranz Bachiller model[28,47] were developed in Crohn's disease patients following ileocolonic resection, the Nakarai model[36] was developed for patients with low (<3mg/L) CRP levels only, and the Beigel first-TNF model[41] selected patients' ileocolonoscopies in follow-up during anti-TNF-α therapy (Table 1 & 2). Therefore, we validated the Nakarai model[36] in patients with a low CRP (TAILORIX: <3mg/L and UAI: <7mg/L) and the Beigel model[41] using the second and third ileocolonoscopies in the TAILORIX dataset. In line with all identified models the most often included predictors in the validated models were CRP (5/7 [71.4%]), fecal calprotectin (5/7 [71.4%]) and symptom-based variables (5/7 [71.4%]). Only for one[20] of the validated models an AUC had been reported in the original publication.

External validation

TAILORIX and UAI baseline characteristics

For external validation we included 155 patients undergoing 346 ileocolonoscopies from the TAILORIX and 82 patients undergoing 93 ileocolonoscopies from the UAI study. The majority of patients were female (Supplementary Material 13). The patients in the UAI development dataset were older, had a longer disease duration, and had less often ileal disease (L1) than patients in the TAILORIX dataset. At week 0 the patients in the TAILORIX dataset had more active disease based on the CDAI, CRP, fecal calprotectin and CDEIS compared to later timepoints in this study and the UAI dataset. The proportion of ileocolonoscopies with endoscopic activity (CDEIS≥3) was 62.1% in the TAILORIX dataset and 54.8% in the UAI dataset.

Discrimination

The AUC-values varied widely between models from 0.58 [95%-CI: 0.39-0.76] (Nakarai model[36] in the UAI dataset) to 0.82 [95%-CI: 0.73-0.91] (Herranz Bachiller model[47] in the UAI dataset) (Figure 2 & Supplementary Material 14). All four models (Bodelier, Herranz Bachiller, Lobaton and Minderhoud (UAI-model)[20,32,43,47]) with an AUC >0.70 used fecal calprotectin, and all of these but one (Herranz Bachiller[47]) used CRP as a predictor. As a benchmark, we therefore also assessed the discriminative performance of fecal calprotectin and CRP as continuous variables and found AUCs >0.70 in both datasets. When we applied generally accepted thresholds[9] for fecal calprotectin (>250 and >100 μg/g) and CRP (>5 in TAILORIX and >7mg/L in UAI) the AUCs were lower, but remained >0.70, except for CRP in the TAILORIX dataset (Figure 2 & Supplementary Material 14).

Calibration and re-calibration

From the validated models, four models[28,32,36,41] validated in the TAILORIX dataset and four models[28,32,36,47] validated in the UAI dataset have as outcome a predicted probability for endoscopic activity. The Minderhoud (UAI) model[20] validated in the TAILORIX dataset predicted the CDEIS as continuous outcome. Logistic recalibration improved the models' calibration in both datasets (Supplementary Materials 15 and 16). The slope corrections for models yielding predicted probabilities ranged from 0.37 to 0.53 in the TAILORIX and 0.19 to 0.47 in the UAI dataset, indicating overestimation of the probabilities for endoscopic activity of the original models. The Minderhoud (UAI) model's[20] predictor weights were less over-fit with a slope correction of 0.81. The model formulas after re-calibration-in-the-large and logistic or linear recalibration are displayed in Supplementary Material 17.

Explained variance in CDEIS by predicted CDEIS, fecal calprotectin and CRP

The marginal R^2 was higher for the Minderhoud (UAI) model[20] (0.40) compared to the continuous fecal calprotectin (0.31) and CRP (0.29) levels in the TAILORIX dataset (Supplementary Material 18).

Avoiding ileocolonoscopies

We sought to identify thresholds in the predicted probabilities, predicted CDEIS, and continuous fecal calprotectin and CRP values in order to potentially avoid ileocolonoscopies for expected EH (based on NPV) or endoscopic activity (based on PPV). A threshold yielding an NPV ≥90% was only found in two models (Garcia-Planella and Herranz Bachiller)[28,47] and not for the individual fecal calprotectin or CRP levels when applied to the UAI dataset. For none of the models and the individual fecal calprotectin and CRP levels such a threshold could be established, when validated in the TAILORIX dataset (Table 3 and 4). In the TAILORIX dataset, thresholds could be identified for a PPV>90% for the Lobaton[32] and Minderhoud (UAI)[20] and a PPV>80% for the Garcia-Planella[28] model. In the UAI dataset a PPV>90% could be reached with thresholds for the Lobaton[32], Herranz-Bachiller[47] and for the Garcia-Planella[28] model. A threshold for a PPV>90% for the individual continuous fecal calprotectin and CRP levels could be identified in both datasets. The models of Beigel[41], Bodelier[43], and Nakarai[36], did not achieve a NPV or PPV ≥80% for any given threshold. Despite not reaching an NPV or PPV ≥80% the Bodelier model had the highest overall accuracy in both validation cohorts compared all other models, fecal calprotectin and CRP (Table 4).

Most correctly avoided ileocolonoscopies are thus in patients with endoscopic activity. Applying the models to the TAILORIX dataset could at best lead to the avoidance of 39.8% [95%-CI: 31.7-48.5%] of all ileocolonoscopies, all for expected endoscopic activity, based on the Lobaton model[32], but 9.8% [95%-CI: 5.3-17.5%] of those ileocolonoscopies would incorrectly be classified as having endoscopic activity. Applying the models to the UAI dataset could lead to avoidance of 41.3% [95%-CI: 26.9-57.4%] of all ileocolonoscopies for both EH and endoscopic activity based on the Herranz-Bachiller model[47], but 10.8% [95%-CI: 4.0-26.1%] of these ileocolonoscopies would be wrongly avoided (Table 3). When applying known thresholds[9] for fecal calprotectin, (<100µg/g and >250µg/g), the proportion of all avoided ileocolonoscopies was larger than with any of the models (TAILORIX: 85.2% [95%-CI: 80.2-89.1%], UAI: 92.4% [95%-CI: 84.7-96.4%]), however this also included the largest proportion of incorrectly avoided ileocolonoscopies (TAILORIX: 21.9% [95%-CI: 17.1-27.8%], UAI: 21.4% [95%-CI: 13.8-31.8%] of all avoided ileocolonoscopies).

DISCUSSION

In our comprehensive systematic review, we identified 21 studies reporting on 27 models for endoscopic activity or EH with an overall high risk of bias in model development. The AUCs of the 7 validated models, consisting of generally available predictors, validated in two separate validation datasets ranged from 0.58-0.82. The best performing models were comparable in their discriminatory abilities with continuous fecal calprotectin levels and to a lesser extent to CRP levels as individual biomarkers. Most models, fecal calprotectin, and CRP could not reach an NPV ≥ 80% for any given threshold indicating that these indexes and biomarkers cannot replace ileocolonoscopies for the monitoring of EH in Crohn's disease. Whether these models render ileocolonoscopies redundant in case of predicted endoscopic activity is a matter of discussion.

CRP and fecal calprotectin are the most frequently used predictors in published prediction models. Of note, we found a comparable performance of prediction models and individual fecal calprotectin levels with an AUC in the TAILORIX of 0.79 [95%-CI: 0.74-0.85], and UAI dataset of 0.82 [95%-CI: 0.73-0.92] when fecal calprotectin was used as continuous predictor. Fecal calprotectin levels were found to only partially explain the variance in CDEIS, however (marginal R^2 in the TAILORIX: 0.31, and UAI dataset: 0.40). CRP as individual biomarker had a lower AUC and marginal R^2 than fecal calprotectin and the best performing models. Although both CRP and fecal calprotectin are established markers of disease activity in CD, there is a considerable inter-individual and intra-individual variation with respect to the magnitude of their response. In case of CRP, this can partly be attributed to polymorphisms in the CRP gene.[50] Furthermore, these biomarkers are affected by the location of the disease[9] and fecal calprotectin may vary considerably from day to day.[51,52]

We did not identify thresholds for an NPV ≥80% for CRP and fecal calprotectin, when applied as continuous values. In other words, endoscopic healing could not reliably enough be predicted. Nonetheless, avoiding colonoscopies based on fecal calprotectin thresholds[9] (<100 μg/g: expected EH and >250 μg/g: expected endoscopic activity) led to the highest proportion of avoided colonoscopies for a correct diagnosis (TAILORIX: 66.5% [95%-CI: 60.8-71.8%], UAI: 72.6% [95%-CI: 62.3-80.9%]). However, this came at the cost of the highest proportion of incorrectly avoided colonoscopies (TAILORIX: 18.7% [95%-CI: 14.5-23.8%], UAI: 19.8% [95%-CI: 12.7-29.6%] of all colonoscopies). Avoiding colonoscopies based on fecal calprotectin thresholds alone can thus not be advocated, because this would lead to undertreatment for incorrect expected EH and unneeded therapy escalation for incorrect expected endoscopic activity in too many patients.

We conclude that published prediction models can presently not be used to replace ileocolonoscopies. The question arises as to how to improve the performance of these tests. The AUCs of four models, i.e. Bodelier[43], Herranz Bachiller[47], Lobaton[32], and Minderhoud (UAI)[20], was higher than fecal calprotectin >250μg/g in one or both datasets. All of these models incorporated fecal calprotectin and CRP or ESR as dichotomized or continuous predictors. It could therefore be that the model performance would have been better if fecal calprotectin and CRP would have been modeled in accordance to their relation with endoscopic activity, e.g. non-linear. Another strategy could be to use temporal changes in biomarker levels, potentially correcting for the inter-individual heterogeneity. Whether other single biomarkers or a combination of biomarkers can replace endoscopy remains to be shown. The Monitr serum panel of 13 biomarkers was recently found to have a NPV of 92% for endoscopic activity[53]. However, in another cohort this serum-based panel did not outperform fecal calprotectin regarding discrimination.[54] Unfortunately, full reports for this serum-based panel have not been published yet.

To increase the chances of developing models that can actually replace ileocolonoscopies in the future, it is of paramount importance that key steps for the development and validation of reliable and reproducible prediction models are undertaken, such as: standard transparent reporting of development strategies and performance criteria (discrimination, calibration, diagnostic accuracy measures)[55], external validation[10], and benchmarking against clinically available and accepted biomarkers (e.g. fecal calprotectin and CRP) as well as adherence to international guidelines, e.g. the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)-statement[55].

To the best of our knowledge, this is the first systematic review and external validation of non-invasive prediction models for endoscopic activity in Crohn's disease. External validation of the identified models in two separate prospective cohorts, with one being the largest in literature currently, enabled us to simultaneously test the performance of all models in two different settings, namely a controlled setting evaluating treatment response with pre-scheduled ileocolonoscopies (TAILORIX-dataset)[19] and consecutive patients with a clinical indication for ileocolonoscopy (UAI-dataset)[20].

Nevertheless, several potential limitations of our study should be considered. Missing predictor values were present in both datasets. This was accounted for in the most appropriate way by multiple imputation.[56] In our external validation, multiple ileocolonoscopies per patient were included, which could have led to biased results. Therefore, we analyzed all performance measures, i.e. discrimination and (re)calibration, accounting for clustering within patients of ileocolonoscopies. Unfortunately, not all identified models could be validated. This was mostly due to poor reporting of the model specifications, or inclusion of investigational biomarkers as predictor. Nonetheless, we were able to validate 7 models with commonly available and easy-to-use non-invasive predictors. The exact definition and thresholds of EH remain a matter of debate. Although a cut-off level of <3 has been proposed for the CDEIS to signify EH, there is no consensus.[3,57] We chose to validate the identified models versus the CDEIS, independent of the endoscopic score used in the individual models. We cannot exclude the possibility that the performances of models would have been different, employing the SES-CD as outcome definition. However, the CDEIS and SES-CD have been shown to have comparably good inter-rater reliability[12], and have been shown to correlate well (correlation coefficient = 0.92) with each other.[58] Therefore, our results based on the CDEIS are likely to be directly reflective of prediction of mucosal activity assessed by the SES-CD.

CONCLUSION

Through a comprehensive systematic review of the literature we were able to externally validate 7 prediction models for endoscopic activity in Crohn's disease in two prospective cohorts. Based on the discriminatory abilities, published prediction models only showed limited to no benefit over fecal calprotectin and CRP as single biomarkers. Avoiding ileocolonoscopies solely based on published prediction models or individual biomarkers seems not yet justified for clinical application, especially due to insufficient certainty in predicting endoscopic healing. Therefore, ileocolonoscopy remains the mainstay for assessment of mucosal activity in Crohn's disease.

FIGURE LEGENDS

Figure 1. Flowchart detailing the number of studies excluded and included at each step of the literature search, and the number of models that could be externally validated within the TAILORIX and UAI dataset. Ten models could not be validated because TAILORIX and UAI dataset do not contain the variables needed for validation: six models included investigational biomarkers, and 4 models included variables not available in the TAILORIX and UAI dataset (Supplementary material 3).

IBD, inflammatory bowel disease; n, number of studies; UAI, Utrecht Activity Index; TAILORIX, a randomized controlled trial investigating tailored treatment with infliximab for active luminal Crohn’s disease.

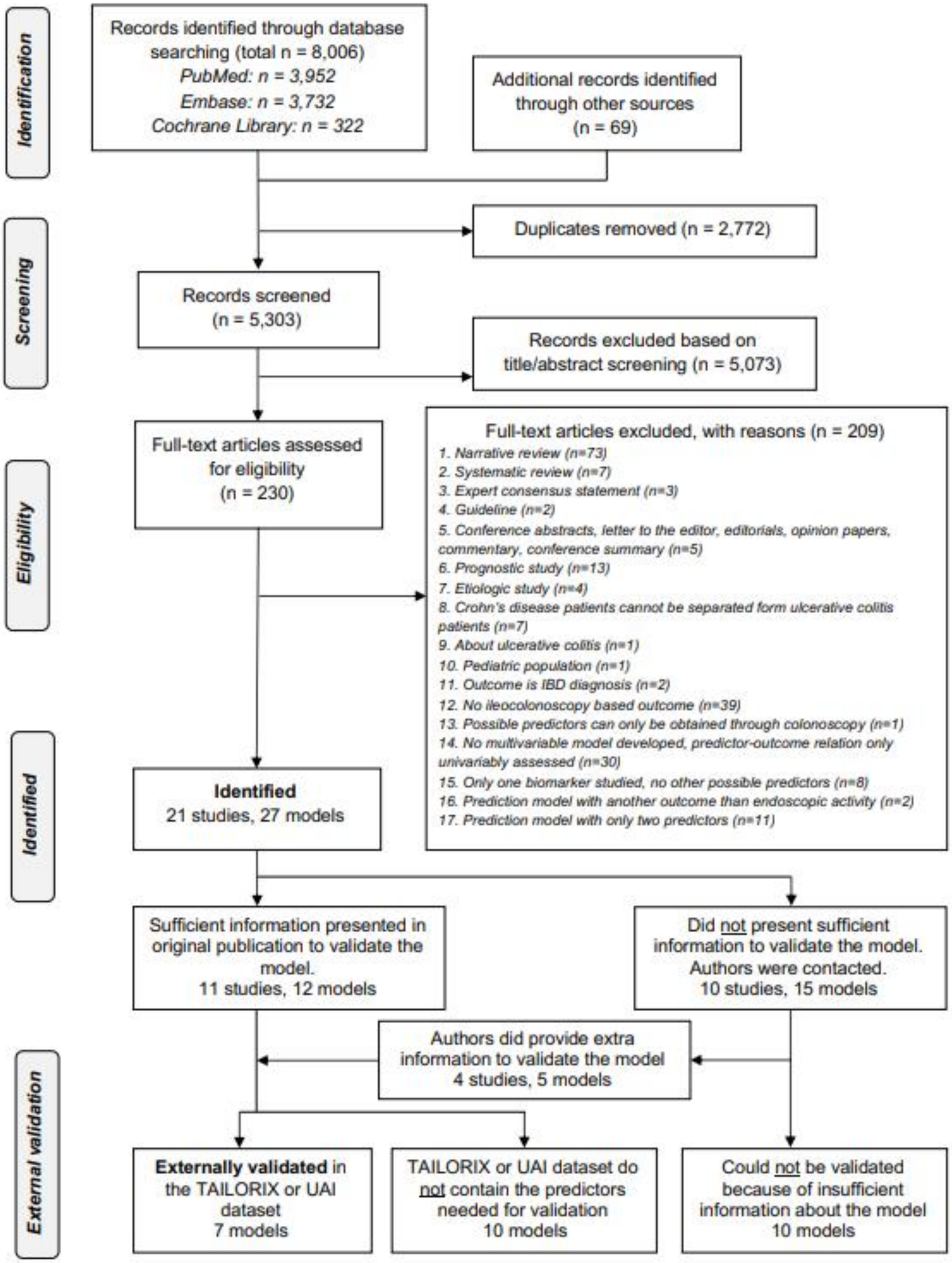
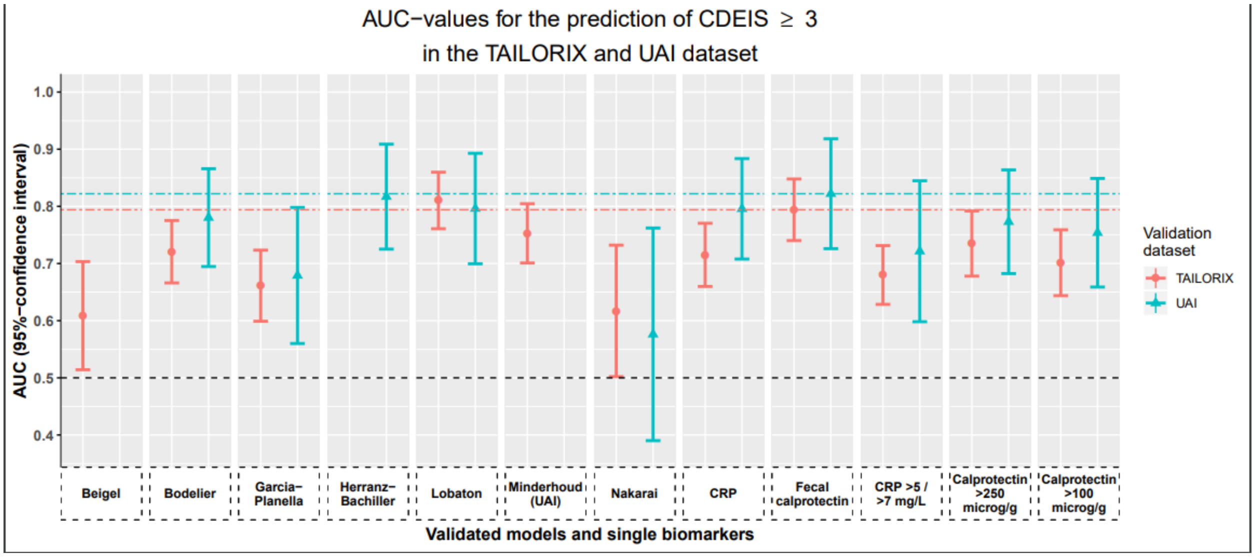

Identification
Records identified through database searching (total n = 8,006)
PubMed: n = 3,952
Embase: n = 3,732
Cochrane Library: n = 322
Additional records identified through other sources (n = 69)
Screening
Duplicates removed (n = 2,772)
Records screened (n = 5,303)
Records excluded based on title/abstract screening (n = 5,073)
Eligibility
Full-text articles assessed for eligibility (n = 230)
Full-text articles excluded, with reasons (n = 209)
1. Narrative review (n=73)
2. Systematic review (n=7)
3. Expert consensus statement (n=3)
4. Guideline (n=2)
5. Conference abstracts, letter to the editor, editorials, opinion papers, commentary, conference summary (n=5)
6. Prognostic study (n=13)
7. Etiologic study (n=4)
8. Crohn's disease patients cannot be separated form ulcerative colitis patients (n=7)
9. About ulcerative colitis (n=1)
10. Pediatric population (n=1)
11. Outcome is IBD diagnosis (n=2)
12. No ileocolonoscopy based outcome (n=39)
13. Possible predictors can only be obtained through colonoscopy (n=1)
14. No multivariable model developed, predictor-outcome relation only univariably assessed (n=30)
15. Only one biomarker studied, no other possible predictors (n=8)
16. Prediction model with another outcome than endoscopic activity (n=2)
17. Prediction model with only two predictors (n=11)
Identified
Identified
21 studies, 27 models
Sufficient information presented in original publication to validate the model.
11 studies, 12 models
Did not present sufficient information to validate the model. Authors were contacted.
10 studies, 15 models
Authors did provide extra information to validate the model
4 studies, 5 models
External validation
Externally validated in the TAILORIX or UAI dataset
7 models
TAILORIX or UAI dataset do not contain the predictors needed for validation
10 models
Could not be validated because of insufficient information about the model
10 models

Figure 2. Discriminative ability of published prediction models, fecal calprotectin and CRP for the outcome endoscopic activity (CDEIS ≥ 3) as tested in the TAILORIX and UAI development dataset. If no AUC is indicated for a model, it was not validated in that particular dataset, because the predictors were not available. The model of Beigel[41] (Beigel-TNF1 model) was only validated within the ileocolonoscopies performed after the baseline colonoscopy in the TAILORIX dataset (191 ileocolonoscopies in 111 patients). The model of Minderhoud (UAI)[20] was not validated in the UAI development dataset, because it was developed in that dataset. The model of Nakarai[36] was only developed and thus validated for patients with a low CRP value (TAILORIX: 115 ileocolonoscopies in 76 patients, UAI: 51 ileocolonoscopies in 49 patients).

The dashed colored lines represent reference lines at the AUC for fecal calprotectin as continuous biomarker in the TAILORIX and UAI development dataset.

AUC, area under the receiver operating characteristic curve; CDEIS, Crohn's disease endoscopic index of severity; CRP, C-reactive protein, TAILORIX, a randomized controlled trial investigating tailored treatment with infliximab for active luminal Crohn's disease; TNF, tumor necrosis factor, UAI, Utrecht Activity Index development dataset.



REFERENCES

1. Pineton de Chambrun G, Blanc P, Peyrin-Biroulet L. Current evidence supporting mucosal healing and deep remission as important treatment goals for inflammatory bowel disease. Expert Rev Gastroenterol Hepatol. 2016;10(8):915-927. doi:10.1586/17474124.2016.1174064.

2. Romkens TEH, Gijsbers K, Kievit W, et al. Treatment Targets in Inflammatory Bowel Disease: Current Status in Daily Practice. J Gastrointestin Liver Dis. 2016;25(4):465-471.

3. Neurath MF, Travis SPL. Mucosal healing in inflammatory bowel diseases: a systematic review. Gut. 2012;61(11):1619-1635. doi:10.1136/gutjnl-2012-302830.

4. Reinink AR, Lee TC, Higgins PDR. Endoscopic Mucosal Healing Predicts Favorable Clinical Outcomes in Inflammatory Bowel Disease: A Meta-analysis. Inflamm Bowel Dis. 2016;22(8):1859-1869. doi:10.1097/MIB.0000000000000816.

5. Best WR, Becktel JM, Singleton JW, et al. Development of a Crohn’s disease activity index. National Cooperative Crohn’s Disease Study. Gastroenterology. 1976;70(3):439-444.

6. Harvey RF, Bradshaw JM. A simple index of Crohn’s-disease activity. Lancet (London,

England). 1980;1(8167):514.

7. Falvey JD, Hoskin T, Meijer B, et al. Disease activity assessment in IBD: clinical indices and biomarkers fail to predict endoscopic remission. Inflamm Bowel Dis. 2015;21(4):824-831. doi:10.1097/MIB.0000000000000341.

8. Peyrin-Biroulet L, Reinisch W, Colombel J-F, et al. Clinical disease activity, C-reactive protein normalisation and mucosal healing in Crohn's disease in the SONIC trial. Gut. 2014;63(1):88-95. doi:10.1136/gutjnl-2013-304984.

9. Mosli MH, Zou G, Garg SK, et al. C-Reactive Protein, Fecal Calprotectin, and Stool Lactoferrin for Detection of Endoscopic Activity in Symptomatic Inflammatory Bowel Disease Patients: A Systematic Review and Meta-Analysis. Am J Gastroenterol. 2015;110(6):802-819; quiz 820. doi:10.1038/ajg.2015.120.

10. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56(9):826-832.

11. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744. doi:10.1371/journal.pmed.1001744.

12. Khanna R, Nelson SA, Feagan BG, et al. Endoscopic scoring indices for evaluation of disease activity in Crohn's disease. Cochrane database Syst Rev. 2016;(8):CD010642. doi:10.1002/14651858.CD010642.pub2.

13. Geersing G-J, Bouwmeester W, Zuithoff P, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7(2):e32844. doi:10.1371/journal.pone.0032844.

14. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8(4):391-397.

15. Mary JY, Modigliani R. Development and validation of an endoscopic index of the severity for Crohn's disease: a prospective multicentre study. Groupe d'Etudes Therapeutiques

des Affections Inflammatoires du Tube Digestif (GETAID). Gut. 1989;30(7):983-989.

16. Daperno M, D'Haens G, Van Assche G, et al. Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD. Gastrointest Endosc. 2004;60(4):505-512.

17. Rutgeerts P, Geboes K, Vantrappen G, et al. Predictability of the postoperative course of Crohn's disease. Gastroenterology. 1990;99(4):956-963.

18. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019;170(1):W1-W33. doi:10.7326/M18-1377.

19. D'Haens G, Vermeire S, Lambrecht G, et al. Increasing Infliximab Dose Based on Symptoms, Biomarkers, and Serum Drug Concentrations Does Not Increase Clinical, Endoscopic, or Corticosteroid-Free Remission in Patients With Active Luminal Crohn's Disease. Gastroenterology. 2018;154(4):1343-1351.e1. doi:10.1053/j.gastro.2018.01.004.

20. Minderhoud IM, Steyerberg EW, van Bodegraven AA, et al. Predicting Endoscopic Disease Activity in Crohn's Disease: A New and Validated Noninvasive Disease Activity Index (The Utrecht Activity Index). Inflamm Bowel Dis. 2015;21(10):2453-2459. doi:10.1097/MIB.0000000000000507.

21. Hebuterne X, Lemann M, Bouhnik Y, et al. Endoscopic improvement of mucosal lesions in patients with moderate to severe ileocolonic Crohn's disease following treatment with certolizumab pegol. Gut. 2013;62(2):201-208. doi:10.1136/gutjnl-2012-302262.

22. Vuitton L, Marteau P, Sandborn WJ, et al. IOIBD technical review on endoscopic indices for Crohn's disease clinical trials. Gut. 2016;65(9):1447-1455. doi:10.1136/gutjnl-2015-309903.

23. R Core Team (2018). R: A language and environment for statistical computing computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-

project.org/.

24. Von Hippel PT. How to imute interactions, squares, and other transformed variables. Sociol Methodol. 2009;39(1):265-291. doi:10.1111/j.1467-9531.2009.01215.x.

25. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1-67.

26. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987. doi:10.2307/3172772.

27. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics. 1997;53(2):567-578.

28. Garcia-Planella E, Manosa M, Cabre E, et al. Fecal Calprotectin Levels Are Closely Correlated with the Absence of Relevant Mucosal Lesions in Postoperative Crohn's Disease. Inflamm Bowel Dis. 2016;22(12):2879-2885. doi:10.1097/MIB.0000000000000960.

29. Steyerberg EW. Chapter 20. Updating for a New Setting. In: Steyerberg EW, ed. Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009:361-366.

30. Lott A, Reiter JP. Wilson Confidence Intervals for Binomial Proportions With Multiple Imputation for Missing Data. Am Stat. May 2018:1-7. doi:10.1080/00031305.2018.1473796.

31. Langhorst J, Elsenbruch S, Koelzer J, et al. Noninvasive markers in the assessment of intestinal inflammation in inflammatory bowel diseases: performance of fecal lactoferrin, calprotectin, and PMN-elastase, CRP, and clinical indices. Am J Gastroenterol. 2008;103(1):162-169. doi:10.1111/j.1572-0241.2007.01556.x.

32. Lobaton T, Lopez-Garcia A, Rodriguez-Moranta F, et al. A new rapid test for fecal calprotectin predicts endoscopic remission and postoperative recurrence in Crohn's disease. J Crohns Colitis. 2013;7(12):e641-51. doi:10.1016/j.crohns.2013.05.005.

33. Ma C, Lumb R, Walker E V, et al. Noninvasive Fecal Immunochemical Testing and Fecal Calprotectin Predict Mucosal Healing in Inflammatory Bowel Disease: A Prospective Cohort Study. Inflamm Bowel Dis. 2017;23(9):1643-1649. doi:10.1097/MIB.0000000000001173.

34. Meuwis M-A, Vernier-Massouille G, Grimaud JC, et al. Serum calprotectin as a biomarker for Crohn's disease. J Crohns Colitis. 2013;7(12):e678-83. doi:10.1016/j.crohns.2013.06.008.

35. Morris MW, Stewart SA, Heisler C, et al. Biomarker-Based Models Outperform Patient-Reported Scores in Predicting Endoscopic Inflammatory Disease Activity. Inflamm Bowel Dis. 2018;24(2):277-285. doi:10.1093/ibd/izx018.

36. Nakarai A, Kato J, Hiraoka S, et al. Slight increases in the disease activity index and platelet count imply the presence of active intestinal lesions in C-reactive protein-negative Crohn's disease patients. Intern Med. 2014;53(17):1905-1911.

37. Nancey S, Boschetti G, Moussata D, et al. Neopterin is a novel reliable fecal marker as accurate as calprotectin for predicting endoscopic disease activity in patients with inflammatory bowel diseases. Inflamm Bowel Dis. 2013;19(5):1043-1052. doi:10.1097/MIB.0b013e3182807577.

38. Viscido A, Corrao G, Taddei G, et al. "Crohn's disease activity index" is inaccurate to detect the post-operative recurrence in Crohn's disease. A GISC study. Ital J Gastroenterol Hepatol. 1999;31(4):274-279. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L29312729.

39. Walters TD, Steinhart AH, Bernstein CN, et al. Validating Crohn's disease activity indices for use in assessing postoperative recurrence. Inflamm Bowel Dis. 2011;17(7):1547-1556. doi:10.1002/ibd.21524.

40. Yarur AJ, Quintero MA, Jain A, et al. Serum Amyloid A as a Surrogate Marker for

Mucosal and Histologic Inflammation in Patients with Crohn’s Disease. Inflamm Bowel Dis. 2017;23(1):158-164. doi:10.1097/MIB.0000000000000991.

41. Beigel F, Deml M, Schnitzler F, et al. Rate and predictors of mucosal healing in patients with inflammatory bowel disease treated with anti-TNF-alpha antibodies. PLoS One. 2014;9(6):e99293. doi:10.1371/journal.pone.0099293.

42. Zittan E, Kabakchiev B, Kelly OB, et al. Development of the Harvey-Bradshaw Index-pro (HBI-PRO) Score to Assess Endoscopic Disease Activity in Crohn’s Disease. J Crohns Colitis. 2017;11(5):543-548. doi:10.1093/ecco-jcc/jjw200.

43. Bodelier AGL, Jonkers D, van den Heuvel T, et al. High Percentage of IBD Patients with Indefinite Fecal Calprotectin Levels: Additional Value of a Combination Score. Dig Dis Sci. 2017;62(2):465-472. doi:10.1007/s10620-016-4397-6.

44. Cellier C, Sahmoud T, Froguel E, et al. Correlations between clinical activity, endoscopic severity, and biological parameters in colonic or ileocolonic Crohn’s disease. A prospective multicentre study of 121 cases. Gut. 1994;35(2):231-235. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L24092870.

45. de Bruyn M, Arijs I, De Hertogh G, et al. Serum Neutrophil Gelatinase B-associated Lipocalin and Matrix Metalloproteinase-9 Complex as a Surrogate Marker for Mucosal Healing in Patients with Crohn’s Disease. J Crohns Colitis. 2015;9(12):1079-1087. doi:10.1093/ecco-jcc/jjv148.

46. Faubion WAJ, Fletcher JG, O’Byrne S, et al. EMerging BiomARKers in Inflammatory Bowel Disease (EMBARK) study identifies fecal calprotectin, serum MMP9, and serum IL-22 as a novel combination of biomarkers for Crohn’s disease activity: role of cross-sectional imaging. Am J Gastroenterol. 2013;108(12):1891-1900. doi:10.1038/ajg.2013.354.

47. Herranz Bachiller MT, Barrio Andres J, Fernandez Salazar L, et al. The utility of faecal

calprotectin to predict post-operative recurrence in Crohns disease. Scand J Gastroenterol. 2016;51(6):720-726. doi:10.3109/00365521.2015.1130164.

48. Jones J, Loftus EVJ, Panaccione R, et al. Relationships between disease activity and serum and fecal biomarkers in patients with Crohn's disease. Clin Gastroenterol Hepatol. 2008;6(11):1218-1224. doi:10.1016/j.cgh.2008.06.010.

49. Klimczak K, Lykowska-Szuber L, Eder P, et al. The diagnostic usefulness of fecal lactoferrin in the assessment of Crohn's disease activity. Eur J Intern Med. 2015;26(8):623-627. doi:10.1016/j.ejim.2015.06.015.

50. Moran CJ, Kaplan JL, Winter HS. Genetic Variation Affects C-Reactive Protein Elevations in Crohn's Disease. Inflamm Bowel Dis. April 2018. doi:10.1093/ibd/izy100.

51. Du L, Foshaug R, Huang VW, et al. Within-Stool and Within-Day Sample Variability of Fecal Calprotectin in Patients With Inflammatory Bowel Disease: A Prospective Observational Study. J Clin Gastroenterol. 2018;52(3):235-240. doi:10.1097/MCG.0000000000000776.

52. Cremer A, Ku J, Amininejad L, et al. Variability of Faecal Calprotectin in Inflammatory Bowel Disease patients: an Observational Case-Control Study. J Crohns Colitis. April 2019. doi:10.1093/ecco-jcc/jjz069.

53. Sandborn WJ, Abreu MT, Dubinsky MC. A Noninvasive Method to Assess Mucosal Healing in Patients* With Crohn's Disease. Gastroenterol Hepatol (N Y). 2018;14(5 Suppl 2):1-12.

54. Battat R, Boland B, Singh S, et al. DIAGNOSTIC ACCURACY OF THE SERUM-BASED MUCOSAL HEALING INDEX ASSAY IN CROHN'S DISEASE AND COMPARATIVE ACCURACY TO FECAL CALPROTECTIN IN ROUTINE PRACTICE. Gastroenterology. 2019;156(3):S20. doi:10.1053/j.gastro.2019.01.078 LK - http://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=15280012&id=doi:10.1053%2Fj.gastro.2019.01.078&atitle=DIAGNOSTIC+ACCURACY+OF+THE+SERUM-

BASED+MUCOSAL+HEALING+INDEX+ASSAY+IN+CROHN%27S+DISEASE+AND+COMPARATIVE+ACCURACY+TO+FECAL+CALPROTECTIN+IN+ROUTINE+PRACTICE&stitle=Gastroenterology&title=Gastroenterology&volume=156&issue=3&spage=S20&epage=&aulast=Battat&aufirst=Robert&auinit=R.&aufull=Battat+R.&coden=&isbn=&pages=S20-&date=2019&auinit1=R&auinitm=.

55. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63. doi:10.7326/M14-0697.

56. Janssen KJM, Vergouwe Y, Donders ART, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55(5):994-1001. doi:10.1373/clinchem.2008.115345.

57. Bossuyt P, Louis E, Mary J-Y, et al. Defining Endoscopic Remission in Ileocolonic Crohn's Disease: Let's Start from Scratch. J Crohns Colitis. 2018;12(10):1245-1248. doi:10.1093/ecco-jcc/jjy097.

58. Khanna R, Zou G, D'Haens G, et al. Reliability among central readers in the evaluation of endoscopic findings from patients with Crohn's disease. Gut. 2016;65(7):1119-1125. doi:10.1136/gutjnl-2014-308973.

TABLES

Table 1. Summary of study, patient and model characteristics of all identified, validated and not validated studies and models.

	All 27 identified models (21 studies)	**7 validated models (7 studies)**	**20 models that could not be validated (15 studies)**
Study characteristics[a], N (%)	**21 studies**	**7 studies**	**15 studies**
Study design of dataset used			
Cohort	19 (90.5%)	7 (100%)	13 (86.7%)
Randomized clinical trial	2 (9.5%)	0	2 (13.3%)
Data-collection			
Prospective	17 (81.0%)	4 (57.1%)	13 (86.7%)
Retrospective	3 (14.3%)	3 (42.9%)	1 (6.7%)
Unknown	1 (4.8%)	0	1 (6.7%)
Multicenter study	9 (42.9%)	2 (28.6%)	7 (46.7%)
Continent			
Europe	13 (61.9%)	6 (85.7%)	8 (53.3%)
North-America	6 (28.6%)	0	6 (40.0%)
Asia	1 (4.8%)	1 (14.3%)	0
Multicontinental	1 (4.8%)	0	1 (6.7%)
Patient domain, N (%)			
Crohn's disease patients in general	13 (61.9%)	3 (42.9%)	10 (66.7%)
Crohn's disease after ileocecal resection	4 (19.0%)	2 (28.6%)	2 (13.3%)
Crohn's disease patients on a certain treatment	3 (14.3%)	1 (14.3%)	3 (20.0%)
Crohn's disease patients with a low CRP level	1 (4.8%)	1 (14.3%)	0
Model characteristics	**27 models**	**7 models**	**20 models**
Number of ileocolonoscopies included for model development, median (range)	89 (32-157)	93 (50-120)	88 (32-157)
Number of patients with the outcome of the model[b], median (range)	39.5 (16-87)	40 (19-87)	39 (16-68)
Not reported:	N: 2	N: 1	N: 1
Continuous outcome:	N: 5	N: 1	N : 4
Number of predictors in the final model, median (range)	3 (3-12), unknown: N=1	3 (3-11)	3 (3-12) unknown: N=1
Top 3 most used predictors			
C-reactive protein (CRP)	17 (63.0%)	5 (71.4%)	12 (60.0%)
Fecal calprotectin	13 (48.1%)	5 (71.4%)	8 (40.0%)
Harvey-Bradshaw Index	5 (18.5%)	2 (28.6%)	3 (15.0%)
Endoscopic score used[c]			
CDEIS	4 (14.8%)	2 (28.6%)	2 (10.0%)
SES-CD	15 (55.6%)	2 (28.6%)	13 (65.0%)
(Modified) Rutgeerts	9 (33.3%)	3 (42.9%)	6 (30.0%)
Endoscopist judgment (no formal score used)	4 (14.8%)	1 (14.3%)	3 (15.0%)
Endoscopy assessment			
Clinical practice	22 (81.5%)[d]	7 (100%)[d]	15 (75.0%)
Central reader(s)	5 (18.5%)	0	5 (25.0%)
Outcome used in model development			
Continuous	5 (18.5%)	1 (14.3%)	4 (20.0%)
Dichotomous	22 (81.5%)	6 (85.7%)	16 (80.0%)

CDEIS, Crohn's disease endoscopic index of severity; N, number of studies/models; SES-CD, Simple endoscopic score for Crohn's disease.

[a] From one study one model could be validated and one model could not be validated, therefore the number of validated and not validated studies separately add up to 22 instead of 21.

[b] Either endoscopic activity or no endoscopic activity based on the definition of the original study.

[c]The total is >100%, because the outcomes of some models were based on a combination of endoscopic scores.
[d]One of these studies, assessed the accuracy by a central reader (intraclass correlation: 0.86), but used the clinical practice values for the model development.

Table 2. Characteristics of models included for external validation.

Author, year	N of ileo-colono-scopies used in model develop-ment	Domain	Original outcome	Original AUC [95%-CI]	Predictors included in the models																	
					Demo-graphics				Symptoms			Treatment related			Time	Laboratory parameters						
					Age	Age at diagnosis	Sex	Smoking	CDAI	HBI	Number of liquid stools during 1 day	Duration of anti-TNF treatment	Time to start anti-TNF treatment	Azathioprine use	Time between colonoscopies	Hemoglobin	WBC	Platelet count	MPV	CRP	ESR	Fecal calprotectin
Beigel, 2014[41]	120	CD patients on anti-TNF treatment	SES-CD >0	NR	●	●	●	●				●	●		●		●[a]			●[a]		
Bodelier, 2017[43]	50	CD patients in general	SES-CD ≥4	NR						●										●		●
Garcia-Planella, 2016[28]	108	CD patients with ileocolonic resection	Rutgeerts ≥i2	NR				●						●						●		●
Herranz Bachiller, 2016[47]	97	CD patients with ileocolonic resection	Modified Rutgeerts ≥i2b	NR						●											●	●
Lobaton, 2013[32]	89	CD patients in general	CDEIS <3 OR Rutgeerts <i2	NR					●											●		●
Minderhoud, 2015[20]	93	CD patients in general	Predicted CDEIS score	CDEIS ≥3: 0.92 [NR]							●							●	●	●		●
Nakarai, 2014[36]	70	CD patients with CRP <3 mg/L	Ulcerations or areas of erosions	NR					●							●		●				

AUC, area under the receiver operating characteristic curve; CD, Crohn's disease; CDAI, Crohn's disease activity index; CDEIS, Crohn's disease endoscopic index of severity; CI, confidence interval; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; HBI, Harvey-Bradshaw Index; MPV, mean platelet volume; N, number of colonoscopies; NR, not reported; SES-CD, Short endoscopic score for Crohn's Disease; TNF, tumor necrosis factor; WBC, white blood cell count.

[a] In this model the CRP and WBC levels are included measured at start of anti-TNF treatment and during follow-up.

● indicates that the predictor is included in the model.

Table 3. Proportions of patients in which a colonoscopy is correctly or incorrectly avoided and still performed based on model threshold values.

First author, year	Model outcome threshold[a]	Threshold based on cut-off of	Colonoscopy avoided based on correct diagnosis % [95%-CI]	Colonoscopy avoided based on incorrect diagnosis % [95%-CI]	Colonoscopy still performed % [95%-CI]
Applied to TAILORIX dataset (N=346)					
Beigel, 2014[41]	Low: NA High: NA	NA NA	-	-	100%
Bodelier, 2017[43]	Low: NA High: NA	NA NA	-	-	100%
Garcia-Planella, 2016[28]	Low: <42% High: ≥86%	NPV ≥ 80% PPV ≥ 80%	16.5% [12.9-20.9%]	3.7% [2.0-6.6%]	79.8% [75.0-83.9%]
Lobaton, 2013[32]	Low: NA High: ≥91%	NA PPV ≥ 90%	35.9% [30.9-41.2%]	3.9% [2.1-7.3%]	60.2% [54.7-65.4%]
Minderhoud, 2015[20]	Low: NA High: UAI ≥6.1	NA PPV ≥ 90%	30.4% [25.7-35.6%]	3.7% [2.1-6.7%]	65.8% [60.4-70.9%]
Nakarai, 2014[36]	Low: NA High: NA	NA NA	-	-	100%
C-reactive protein	Low: NA High: ≥17 mg/L	NA PPV ≥ 90%	21.9% [17.7-26.7%]	2.4% [1.1-4.9%]	75.8% [70.7-80.2%]
Fecal calprotectin	Low: NA High:≥1283μg/g	NA PPV ≥ 90%	23.0% [18.7-28.0%]	2.5% [1.1-5.2%]	74.5% [69.2-79.2%]
Fecal calprotectin	Low: <100 μg/g High: >250 μg/g	Literature[9]	66.5% [60.8-71.8%]	18.7% [14.5-23.8%]	14.8% [11.2-19.4%]
Applied to Utrecht Activity Index dataset (N=93)					
Bodelier, 2017[43]	Low: NA High: NA	NA NA	-	-	100%
Garcia-Planella, 2016[28]	Low: <36% High: ≥85%	NPV ≥ 90% PPV ≥ 90%	18.5% [11.8-27.7%]	1.3% [0.2-9.3%]	80.2% [70.3-87.4%]
Herranz Bachiller, 2016[47]	Low: <17% High: ≥65%	NPV ≥ 90% PPV ≥ 90%	36.8% [27.7-47.1%]	4.5% [1.6-11.8%]	58.7% [48.3-68.4%]
Lobaton, 2013[32]	Low: NA High: ≥78%	NA PPV ≥ 90%	35.0% [26.0-45.3%]	3.9% [1.4-10.3%]	61.1% [50.8-70.5%]
Nakarai, 2014[36]	Low: NA High: NA	NA NA	-	-	100%
C-reactive protein	Low: NA High: ≥20 mg/L	NA PPV ≥ 90%	17.3% [10.9-26.3%]	1.1% [0.2-5.8%]	81.6% [72.5-88.2%]
Fecal calprotectin	Low: NA High: ≥856 μg/g	NA PPV ≥ 90%	28.0% [19.9-38.0%]	2.0% [0.5-8.0%]	70.0% [59.9-78.4%]
Fecal calprotectin	Low: <100 μg/g High: >250 μg/g	Literature[9]	72.6% [62.3-80.9%]	19.8% [12.7-29.6%]	7.6% [3.7-14.9%]

95%-CI, 95% confidence interval; N, number of colonoscopies; NA, not available; NPV, negative predictive value, PPV, positive predictive value.

[a] The thresholds depict predicted probabilities if percentages are shown, the predicted CDEIS for the Minderhoud model and single biomarker values for fecal calprotectin and CRP. "Low" reflects the threshold based on the NPV indicating expected endoscopic healing and high the threshold based on the PPV indicating expected endoscopic activity.

Table 4. Diagnostic values for identified thresholds per model

First author of the model or biomarker	Model outcome threshold	Threshold based on cut-off of	NPV % [95%-CI]	PPV % [95%-CI]	Sensitivity % [95%-CI]	Specificity % [95%-CI]	Overall accuracy % [95%-CI]
Tailorix dataset (N=346)							
Beigel[41]	Low: NA	NA	-	-	-	-	-
	High: NA	NA	-	-	-	-	-
Bodelier[43,a]	Low: NA	NA	-	-	-	-	-
	High: NA	NA	71.0 [65.7-75.8]	77.3 [72.4-81.5]	85.3 [80.8-88.9]	58.8 [53.1-64.2]	75.3 [70.1-79.8]
Garcia-Planella[28]	Low: <42%	NPV ≥ 80%	83.4 [78.5-87.3]	66.2 [61.1-71.0]	97.8 [95.2-99.0]	18.2 [14.4-22.7]	67.6 [62.5-72.4]
	High: ≥86%	PPV ≥ 80%	40.4 [35.3-45.7]	80.9 [76.1-85.0]	15.5 [12.0-19.8]	94.0 [90.6-96.2]	45.2 [40.0-50.6]
Lobaton[32]	Low: NA	NA	-	-	-	-	-
	High: ≥91%	PPV ≥ 90%	56.4 [51.0-61.7]	90.2 [85.7-93.4]	57.8 [52.3-63.1]	89.7 [85.0-93.0]	69.8 [64.4-74.8]
Minderhoud[20]	Low: NA	NA	-	-	-	-	-
	High: UAI ≥6.1	PPV ≥ 90%	51.8 [46.5-57.1]	89.0 [84.9-92.2]	48.9 [43.5-54.4]	90.1 [86.0-93.1]	64.5 [59.2-69.5]
Nakarai[36]	Low: NA	NA	-	-	-	-	-
	High: NA	NA	-	-	-	-	-
C-reactive protein	Low: NA	NA	-	-	-	-	-
	High:≥17 mg/L	PPV ≥ 90%	46.9 [41.6-52.2]	90.3 [86.3-93.2]	35.2 [30.2-40.6]	93.8 [90.2-96.1]	57.4 [52.0-62.7]
Fecal calprotectin	Low: NA	NA	-	-	-	-	-
	≥1283 µg/g	PPV ≥ 90%	47.5 [42.2-52.8]	90.4 [86.2-93.4]	37.0 [31.8-42.6]	93.5 [89.7-96.0]	58.4 [52.9-63.7]
Fecal calprotectin	<100 µg/g	Literature[9]	74.1 [68.8-78.8]	64.7 [59.5-69.6]	97.2 [94.6-98.6]	13.0 [9.7-17.1]	65.3 [60.1-70.2]
	>250 µg/g	Literature[9]	63.3 [57.7-68.5]	81.7 [76.8-85.5]	74.3 [69.0-78.9]	72.7 [66.9-77.8]	73.7 [68.1-78.6]
Utrecht Activity Index dataset (N=93)							
Bodelier[43,a]	Low: NA	NA	-	-	-	-	-
	High: NA	NA	77.9 [68.3-85.2]	79.0 [69.5-86.1]	82.8 [73.8-89.2]	73.2 [63.3-81.3]	78.5 [68.9-85.7]
Garcia-Planella[28]	Low: <36%	NPV ≥ 90%	94.3 [85.8-97.8]	61.5 [51.3-70.8]	98.7 [92.3-99.8]	25.0 [17.2-34.9]	65.4 [55.2-74.4]
	High: ≥85%	PPV ≥ 90%	48.3 [38.4-58.3]	92.4 [82.8-96.9]	13.1 [7.6-21.6]	98.6 [91.8-99.8]	51.7 [41.6-61.7]
Herranz Bachiller[47]	Low: <17%	NPV ≥ 90%	89.9 [78.9-95.5]	57.1 [46.9-66.7]	99.0 [92.1-99.9]	9.5 [5.0-17.4]	58.6 [48.3-68.2]
	High: ≥65%	PPV ≥ 90%	64.9 [54.6-73.9]	89.2 [80.7-94.2]	59.3 [49.0-68.9]	91.3 [83.2-95.7]	73.7 [63.6-81.9]
Lobaton[32]	Low: NA	NA	-	-	-	-	-
	High: ≥78%	PPV ≥ 90%	67.6 [57.4-76.4]	90.1 [82.0-94.8]	63.9 [53.5-73.2]	91.5 [83.6-95.7]	76.3 [66.4-84.0]
Nakarai[36]	Low: NA	NA	-	-	-	-	-
	High: NA	NA	-	-	-	-	-
C-reactive protein	Low: NA	NA	-	-	-	-	-
	High:≥20 mg/L	PPV ≥ 90%	54.0 [43.9-63.8]	94.1 [87.4-97.4]	31.5 [23.0-41.6]	97.6 [92.2-99.3]	61.4 [51.2-70.7]
Fecal calprotectin	Low: NA	NA	-	-	-	-	-
	≥856 µg/g	PPV ≥ 90%	61.7 [51.5-71.0]	93.5 [85.7-97.2]	51.1 [41.0-61.2]	95.6 [88.5-98.4]	71.2 [61.1-79.6]
Fecal calprotectin	<100 µg/g	Literature[9]	74.3 [64.2-82.3]	77.0 [67.3-84.5]	79.7 [69.9-86.9]	71.0 [60.9-79.4]	75.8 [65.8-83.6]
	>250 µg/g	Literature[9]	71.8 [61.6-80.1]	82.3 [73.1-88.9]	73.9 [63.7-82.0]	80.8 [71.2-87.7]	77.0 [66.9-84.7]

95%-CI, 95% confidence interval; N, number of colonoscopies; NA, not available; NPV, negative predictive value, PPV, positive predictive value; UAI, Utrecht activity index.

[a] The model of Bodelier did not reach the cut-offs for PPV and NPV. Nevertheless, we evaluated its diagnostic accuracy, because this decision rule does not provide a probability but only expected presence or absence of endoscopic activity.