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. Brand1,2, MD, Sjoerd G. Elias3, MD, PhD, Itta M. Minderhoud4, MD, PhD, Julius J. van der Veen1, BSc, LLB, Filip J. Baert5, MD, PhD, David Laharie6, MD, PhD, Peter Bossuyt7, MD, Yoram Bouhnik8, MD, PhD, Anthony Buisson9, MD, Guy Lambrecht10, MD, Edouard Louis11, MD, PhD, Benjamin Pariente12, MD, PhD, Marieke J. Piér13, MD, PhD, C. Janneke van der Woude14, MD, PhD, Geert R.A.M. D’Haens15, MD, PhD, Séverine Vermeire16, MD, PhD, Bas Oldenburg1, 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.
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, Ferring, Janssen, MSD, Pfizer, Takeda, and Lilly.

CJvW received grants from Abbvie, Janssen, 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, Boehringer Ingelheim, Celgene/Receptos, Celltrion, Cosmo, Covidien/Medtronic, Ferring, DrFALK Pharma, Eli Lilly, Engene, Galapagos, Genentech/Roche, Gilead, Glaxo Smith Kline, Hospira/Pfizer, Immune, 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, JJVvdV, 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. Meuvis, 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\textsuperscript{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\textsuperscript{12} (Supplementary Material 2). The modified Ingui filter\textsuperscript{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)\textsuperscript{15}, Simple Endoscopic Scale for Crohn’s Disease (SES-CD)\textsuperscript{16}, Rutgeerts score\textsuperscript{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)\textsuperscript{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.

\textit{Utrecht activity index dataset}

The Utrecht Activity Index (UAI) study aimed to develop and externally validate a model that could predict the CDEIS\textsuperscript{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)\textsuperscript{20}, the CDEIS scores from the first endoscopist were used for the present study.

\textbf{Predictors and outcome in external validation}

\textit{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.\textsuperscript{15} A CDEIS≥3 was considered endoscopic activity\textsuperscript{3}, hence EH was defined as a CDEIS<3.\textsuperscript{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.\textsuperscript{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\textsuperscript{24}) by iterative (25 iterations) chained equations employing the MICE package.\textsuperscript{25} Analyses were performed in all multiple imputed datasets and subsequently pooled by Rubin’s rules.\textsuperscript{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 one28 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 $\geq 90\%$ or, if not attainable, $\geq 80\%$. 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 $\mu$g/g or $>250$ $\mu$g/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 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. 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 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 representing five models. In total, 20 models could not be externally validated, because: 1) ten models did not provide sufficient detail to allow external validation, 2) six models included investigational biomarkers, 3) four models used predictors not available in the TAILORIX and/or UAI dataset (Supplementary material 12). Nonetheless, seven models 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 were developed in Crohn’s disease patients in general, the Garcia-Planella and Herranz Bachiller model were developed in Crohn’s disease patients following ileocolonic resection, the Nakarai model was developed for patients with low (≤3mg/L) CRP levels only, and the Beigel first-TNF model selected patients’ ileocolonoscopies in follow-up during anti-TNF-α therapy (Table 1 & 2). Therefore, we validated the Nakarai model in patients with a low CRP (TAILORIX: <3mg/L and UAI: <7mg/L) and the Beigel model 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 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 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 R² was higher for the Minderhoud (UAI) model²⁰ (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 Lobaton³² and Minderhoud (UAI)²⁰ and a PPV>80% for the Garcia-Planella²⁸ model. In the UAI dataset a PPV>90% could be reached with thresholds for the Lobaton³², Herranz-Bachiller⁴⁷ and for the Garcia-Planella²⁸ 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 Beigel⁴¹, Bodelier⁴³, and Nakarai³⁶, 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\textsuperscript{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\textsuperscript{47}, but 10.8% [95%-CI: 4.0-26.1]% of these ileocolonoscopies would be wrongly avoided (Table 3). When applying known thresholds\textsuperscript{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.\textsuperscript{50} Furthermore, these biomarkers are affected by the location of the disease\textsuperscript{9} and fecal calprotectin may vary considerably from day to day.\textsuperscript{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 \(\mu\)g/g: expected EH and >250 \(\mu\)g/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 >250\(\mu\)g/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)\textsuperscript{55}, external validation\textsuperscript{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\textsuperscript{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)\textsuperscript{19} and consecutive patients with a clinical indication for ileocolonoscopy (UAI-dataset)\textsuperscript{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.\textsuperscript{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.\textsuperscript{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\textsuperscript{12}, and have been shown to correlate well (correlation coefficient = 0.92) with each other.\textsuperscript{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.
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\textsuperscript{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)\textsuperscript{20} was not validated in the UAI development dataset, because it was developed in that dataset. The model of Nakarai\textsuperscript{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


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


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


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


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

prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement.


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

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

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

*aFrom 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.

*bEither endoscopic activity or no endoscopic activity based on the definition of the original study.
The total is >100%, because the outcomes of some models were based on a combination of endoscopic scores. 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.

<table>
<thead>
<tr>
<th>Author, year</th>
<th>N of ileo-colonoscopies used in model development</th>
<th>Domain</th>
<th>Original outcome</th>
<th>Original AUC [95% CI]</th>
<th>Predictors included in the models</th>
<th>Demographics</th>
<th>Symptoms</th>
<th>Treatment related</th>
<th>Time</th>
<th>Laboratory parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Age at diagnosis</td>
<td>Sex</td>
<td>Smoking</td>
<td>CDAI</td>
<td>HBI</td>
<td>Number of liquid stools during 1 day</td>
</tr>
<tr>
<td>Beigel, 2014</td>
<td>120</td>
<td>CD patients on anti-TNF treatment</td>
<td>SES-CD &gt;0</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Bodelier, 2017</td>
<td>50</td>
<td>CD patients in general</td>
<td>SES-CD ≥4</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Garcia-Planella, 2016</td>
<td>108</td>
<td>CD patients with ileocolonic resection</td>
<td>Rutgeerts ≥12</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Herranz Bachiller, 2016</td>
<td>97</td>
<td>CD patients with ileocolonic resection</td>
<td>Modified Rutgeerts ≥12b</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Lobaton, 2013</td>
<td>89</td>
<td>CD patients in general</td>
<td>CDEIS &lt;3 OR Rutgeerts &lt;12</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Minderhoud, 2015</td>
<td>93</td>
<td>CD patients in general</td>
<td>Predicted CDEIS score</td>
<td>CDEIS ≥3: 0.92 [NR]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Nakarai, 2014</td>
<td>70</td>
<td>CD patients with CRP &lt;3 mg/L</td>
<td>Ulcerations or areas of erosions</td>
<td>NR</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>

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.

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.

<table>
<thead>
<tr>
<th>First author, year</th>
<th>Model outcome threshold</th>
<th>Threshold based on cut-off of</th>
<th>Colonoscopy avoided based on correct diagnosis % [95%-CI]</th>
<th>Colonoscopy avoided based on incorrect diagnosis % [95%-CI]</th>
<th>Colonoscopy still performed % [95%-CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied to TAILORIX dataset (N=346)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beigel, 201441</td>
<td>Low: NA</td>
<td>NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bodelier, 201743</td>
<td>Low: NA</td>
<td>NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Garcia-Planella, 201628</td>
<td>Low: &lt;42% High: ≥86%</td>
<td>NPV ≥ 80% PPV ≥ 80%</td>
<td>16.5% [12.9-20.9%]</td>
<td>3.7% [2.0-6.6%]</td>
<td>79.8% [75.0-83.9%]</td>
</tr>
<tr>
<td>Lobaton, 201332</td>
<td>Low: NA High: ≥91%</td>
<td>NA PPV ≥ 90%</td>
<td>35.9% [30.9-41.2%]</td>
<td>3.9% [2.1-7.3%]</td>
<td>60.2% [54.7-65.4%]</td>
</tr>
<tr>
<td>Minderhoud, 201520</td>
<td>Low: NA High: UAI ≥6.1</td>
<td>NA PPV ≥ 90%</td>
<td>30.4% [25.7-35.6%]</td>
<td>3.7% [2.1-6.7%]</td>
<td>65.8% [60.4-70.9%]</td>
</tr>
<tr>
<td>Nakarai, 201436</td>
<td>Low: NA High: NA</td>
<td>NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>Low: NA High: ≥17 mg/L</td>
<td>NA PPV ≥ 90%</td>
<td>21.9% [17.7-26.7%]</td>
<td>2.4% [1.1-4.9%]</td>
<td>75.8% [70.7-80.2%]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>Low: NA High:≥1283 μg/g</td>
<td>NA PPV ≥ 90%</td>
<td>23.0% [18.7-28.0%]</td>
<td>2.5% [1.1-5.2%]</td>
<td>74.5% [69.2-79.2%]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>Low: &lt;100 μg/g High: &gt;250 μg/g</td>
<td>Literature9</td>
<td>66.5% [60.8-71.8%]</td>
<td>18.7% [14.5-23.8%]</td>
<td>14.8% [11.2-19.4%]</td>
</tr>
</tbody>
</table>

Applied to Utrecht Activity Index dataset (N=93)

| Bodelier, 201743 | Low: NA High: NA | NA | NA | - | - | 100% |
| Garcia-Planella, 201628 | 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, 201647 | 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, 201332 | 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, 201436 | 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 | Literature9 | 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.

aThe 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

<table>
<thead>
<tr>
<th>First author of the model or biomarker</th>
<th>Model outcome threshold</th>
<th>Threshold based on cut-off of</th>
<th>NPV % [95% CI]</th>
<th>PPV % [95% CI]</th>
<th>Sensitivity % [95% CI]</th>
<th>Specificity % [95% CI]</th>
<th>Overall accuracy % [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tailorix dataset (N=346)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beigel61</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: NA</td>
<td>NA</td>
<td>71.0 [65.7-75.8]</td>
<td>77.3 [72.4-81.5]</td>
<td>85.3 [80.8-88.9]</td>
<td>58.8 [53.1-64.2]</td>
<td>75.3 [70.1-79.8]</td>
</tr>
<tr>
<td>Bodelier43,a</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: NA</td>
<td>NA</td>
<td>71.0 [65.7-75.8]</td>
<td>77.3 [72.4-81.5]</td>
<td>85.3 [80.8-88.9]</td>
<td>58.8 [53.1-64.2]</td>
<td>75.3 [70.1-79.8]</td>
</tr>
<tr>
<td>Garcia-Planella28</td>
<td>Low: &lt;42%</td>
<td>NPV ≥ 80%</td>
<td>83.4 [78.5-87.3]</td>
<td>66.2 [61.1-71.0]</td>
<td>97.8 [95.2-99.0]</td>
<td>18.2 [14.4-22.7]</td>
<td>67.6 [62.5-72.4]</td>
</tr>
<tr>
<td></td>
<td>High: ≥86%</td>
<td>PPV ≥ 80%</td>
<td>40.4 [35.3-45.7]</td>
<td>80.9 [76.1-85.0]</td>
<td>15.5 [12.0-19.8]</td>
<td>94.0 [90.6-96.2]</td>
<td>45.2 [40.0-50.6]</td>
</tr>
<tr>
<td>Lobaton12</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: ≥91%</td>
<td>PPV ≥ 90%</td>
<td>56.4 [51.0-61.7]</td>
<td>90.2 [85.7-93.4]</td>
<td>57.8 [52.3-63.1]</td>
<td>89.7 [85.0-93.0]</td>
<td>69.8 [64.4-74.8]</td>
</tr>
<tr>
<td>Minderhoud20</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: UAI≥6.1</td>
<td>PPV ≥ 90%</td>
<td>51.8 [46.5-57.1]</td>
<td>89.0 [84.9-92.2]</td>
<td>48.9 [43.5-54.4]</td>
<td>90.1 [86.0-93.1]</td>
<td>64.5 [59.2-69.5]</td>
</tr>
<tr>
<td>Nakarai36</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: ≥17 mg/L</td>
<td>PPV ≥ 90%</td>
<td>46.9 [41.6-52.2]</td>
<td>90.3 [86.3-93.2]</td>
<td>35.2 [30.2-40.6]</td>
<td>93.8 [90.2-96.1]</td>
<td>57.4 [52.0-62.7]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>≥1283 µg/g</td>
<td>PPV ≥ 90%</td>
<td>47.5 [42.2-52.8]</td>
<td>90.4 [86.2-93.4]</td>
<td>37.0 [31.8-42.6]</td>
<td>93.5 [89.7-96.0]</td>
<td>58.4 [52.9-63.7]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>&lt;100 µg/g</td>
<td>Literature9</td>
<td>74.1 [68.8-78.8]</td>
<td>64.7 [59.5-69.6]</td>
<td>97.2 [94.6-98.6]</td>
<td>13.0 [9.7-17.1]</td>
<td>65.3 [60.1-70.2]</td>
</tr>
<tr>
<td></td>
<td>&gt;250 µg/g</td>
<td>Literature9</td>
<td>63.3 [57.7-68.5]</td>
<td>81.7 [76.8-85.5]</td>
<td>74.3 [69.0-78.9]</td>
<td>72.7 [66.9-77.8]</td>
<td>73.7 [68.1-78.6]</td>
</tr>
<tr>
<td>Utrecht Activity Index dataset (N=93)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bodelier43,a</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: &lt;36%</td>
<td>NPV ≥ 90%</td>
<td>94.3 [85.8-97.8]</td>
<td>61.5 [51.3-70.8]</td>
<td>98.7 [92.3-99.8]</td>
<td>25.0 [17.2-34.9]</td>
<td>65.4 [55.2-74.4]</td>
</tr>
<tr>
<td>Garcia-Planella28</td>
<td>Low: &lt;36%</td>
<td>NPV ≥ 90%</td>
<td>94.3 [85.8-97.8]</td>
<td>61.5 [51.3-70.8]</td>
<td>98.7 [92.3-99.8]</td>
<td>25.0 [17.2-34.9]</td>
<td>65.4 [55.2-74.4]</td>
</tr>
<tr>
<td>Herranz Bachiller47</td>
<td>Low: &lt;17%</td>
<td>NPV ≥ 90%</td>
<td>89.9 [78.9-95.5]</td>
<td>57.1 [46.9-66.7]</td>
<td>99.0 [92.1-99.9]</td>
<td>9.5 [5.0-17.4]</td>
<td>58.6 [48.3-68.2]</td>
</tr>
<tr>
<td></td>
<td>High: ≥65%</td>
<td>PPV ≥ 90%</td>
<td>64.9 [54.6-73.9]</td>
<td>89.2 [80.7-94.2]</td>
<td>59.3 [49.0-68.9]</td>
<td>91.3 [83.2-95.7]</td>
<td>73.7 [63.6-81.9]</td>
</tr>
<tr>
<td>Lobaton12</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: ≥78%</td>
<td>PPV ≥ 90%</td>
<td>67.6 [57.4-76.4]</td>
<td>90.1 [82.0-94.8]</td>
<td>63.9 [53.5-73.2]</td>
<td>91.5 [83.6-95.7]</td>
<td>76.3 [66.4-84.0]</td>
</tr>
<tr>
<td>Nakarai36</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>High: ≥20 mg/L</td>
<td>PPV ≥ 90%</td>
<td>54.0 [43.9-63.8]</td>
<td>94.1 [87.4-97.4]</td>
<td>31.5 [23.0-41.6]</td>
<td>97.6 [92.2-99.3]</td>
<td>61.4 [51.2-70.7]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>Low: NA</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>≥856 µg/g</td>
<td>PPV ≥ 90%</td>
<td>61.7 [51.5-71.0]</td>
<td>93.5 [85.7-97.2]</td>
<td>51.1 [41.0-61.2]</td>
<td>95.6 [88.5-98.4]</td>
<td>71.2 [61.1-79.6]</td>
</tr>
<tr>
<td>Fecal calprotectin</td>
<td>&lt;100 µg/g</td>
<td>Literature9</td>
<td>74.3 [64.2-82.3]</td>
<td>77.0 [67.3-84.5]</td>
<td>79.7 [69.9-86.9]</td>
<td>71.0 [60.9-79.4]</td>
<td>75.8 [65.8-83.6]</td>
</tr>
<tr>
<td></td>
<td>&gt;250 µg/g</td>
<td>Literature9</td>
<td>71.8 [61.6-80.1]</td>
<td>82.3 [73.1-88.9]</td>
<td>73.9 [63.7-82.0]</td>
<td>80.8 [71.2-87.7]</td>
<td>77.0 [66.9-84.7]</td>
</tr>
</tbody>
</table>

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

*aThe 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.*