Article published online: 2022-07-15

Clinical application and diagnostic accuracy of artificial intelligence in colonoscopy for inflammatory bowel disease: systematic review

OPEN ACCESS

Authors

Linda S. Yang[1], Evelyn Perry[1], Leonard Shan[2], Helen Wilding[3], William Connell[1], Alexander J. Thompson[1], Andrew C. F. Taylor[1], Paul V. Desmond[1], Bronte A. Holt[1]

Institutions

1 Department of Gastroenterology, St. Vincent's Hospital and the University of Melbourne, Fitzroy, Victoria, Australia
2 Department of Surgery, Faculty of Medicine, Dentistry and Health Sciences, the University of Melbourne, Fitzroy, Victoria, Australia
3 Library Service, St. Vincent's Hospital Melbourne, Fitzroy, Victoria, Australia

submitted 30.11.2021
accepted after revision 2.5.2022

Bibliography
Endosc Int Open 2022; 10: E1004–E1013
DOI 10.1055/a-1846-0642
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Corresponding author
Linda S. Yang, Department of Gastroenterology, St. Vincent's Hospital Melbourne Pty Ltd, 35 Victoria Pde, Fitzroy, 3065, VIC, Australia
Fax: +0392313644
lindaviva0912@gmail.com

Supplementary material is available under https://doi.org/10.1055/a-1846-0642

ABSTRACT

Background and aims Artificial intelligence (AI) technology is being evaluated for its potential to improve colonoscopic assessment of inflammatory bowel disease (IBD), particularly with computer-aided image classifiers. This review evaluates the clinical application and diagnostic test accuracy (DTA) of AI algorithms in colonoscopy for IBD.

Methods A systematic review was performed on studies evaluating AI in colonoscopy of adult patients with IBD. MEDLINE, Embase, Emcare, PsycINFO, CINAHL, Cochrane Library and Clinicaltrials.gov databases were searched on 28th April 2021 for English language articles published between January 1, 2000 and April 28, 2021. Risk of bias and applicability were assessed with the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Diagnostic accuracy was presented as median (interquartile range).

Results Of 1029 records screened, nine studies with 7813 patients were included for review. AI was used to predict endoscopic and histologic disease activity in ulcerative colitis, and differentiation of Crohn's disease from Behcet's disease and intestinal tuberculosis. DTA of AI algorithms ranged between 52–91 %. The sensitivity and specificity for AI algorithms predicting endoscopic severity of disease were 78 % (range 72–83, interquartile range 5.5) and 91 % (range 86–96, interquartile range 5), respectively.

Conclusions AI has been primarily used to assess disease activity in ulcerative colitis. The diagnostic performance is promising and suggests potential for other clinical application of AI in IBD colonoscopy such as dysplasia detection. However, current evidence is limited by retrospective data and models trained on still images only. Future prospective multicenter studies with full-motion videos are needed to replicate the real-world clinical setting.

Introduction

Inflammatory bowel disease (IBD) is a chronic relapsing condition affecting more than 6.8 million people worldwide [1]. Colonoscopy is a cornerstone in diagnosis and management. It allows endoscopic characterization and tissue acquisition for histologic examination which are the gold standard in identification of IBD phenotype and severity. This is necessary for therapeutic decision-making and evaluating treatment response.

Furthermore, patients with long-standing IBD colitis are at an increased risk of colorectal cancer [2]. Colonoscopic surveillance is important and increasingly aided by enhanced imaging techniques [3, 4].

Despite being standard of care in IBD, endoscopic evaluation is subjective and limited by interobserver and intra-observer variability [5]. To address this, standardized endoscopic assessment tools for disease severity have been developed. These include the Mayo endoscopic score (MES) and ulcerative colitis endoscopy index of severity (UCEIS) [6] and the simple endoscopic score for Crohn's disease (SES-CD). [7] However, considerable interobserver differences remain in endoscopic evaluation of disease activity [5, 7, 8].

Artificial intelligence (AI) refers to machine learning to replicate or simulate human intelligence and neural networks are the most commonly utilized AI subtype in endoscopy. Image recognition is an important aspect of AI that aids endoscopic detection of pathology. Endoscopic recognition of colonic polyps using AI systems has the largest body of evidence with successful clinical translation and uptake of the technology in real-life settings in many countries [9–11]. AI-based polyp detection systems during colonoscopy have been shown to improve adenoma detection rate and in particular, the detection of small, non-advanced adenomas [12].

Computer-aided image interpretation in IBD endoscopy is in its infancy. Growing evidence demonstrates the potential to minimize interobserver variability in endoscopic evaluation, thus reducing variation in patient care and outcomes. A recent systematic review on this topic offers a narrative overview on the use of AI in various subtypes of gastrointestinal endoscopy, including capsule endoscopy, endomicroscopy and experimental techniques such as prototype endoscope using light emitting diode illumination [13]. A number of narrative reviews have also been published in recent years, highlighting the contemporary interest and clinical need for AI application in IBD. These reviews summarize a range of AI application in IBD, including computer-aided endoscopy and algorithms for risk and treatment response prediction [14–16]. However, there is no systematic review focusing solely on the use of AI on conventional colonoscopy, even though colonoscopy is the most routine and essential endoscopic procedure for IBD diagnosis and management. A quantitative analysis of the diagnostic test accuracy (DTA) of AI models is also lacking in current literature.

Therefore, the aims of this systematic review are:

1. To evaluate the current clinical applications of AI algorithms in IBD colonoscopy, and
2. To analyze the DTA of AI algorithms in IBD colonoscopy.

Methods

This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy (PRISMA-DTA) [17]. The protocol was registered with PROSPERO (Registration number CRD42021252612).

Eligibility criteria

The interventions studied in this systematic review for IBD include AI algorithms for use in colonoscopic still images and videos. Studies reporting the clinical application and diagnostic test accuracy (DTA) of AI in colonoscopy for IBD were included. Strict eligibility criteria were followed (**Supplementary Table 1**). AI algorithms using conventional colonoscopy imaging (white light and virtual or dye-based chromoendoscopy) were included.

Information sources

A comprehensive search was performed on April 28, 2021 in Ovid MEDLINE(R) ALL 1946 to April 26, 2021, Embase 1974 to April 26, 2021 (Ovid), Ovid Emcare 1995 to 2021 Week 16, APA PsychInfo 1806 to April Week 3 2021 (Ovid), CINAHL (EBSCOhost), Cochrane Library (Wiley) and Clinicaltrials.gov.

Search

Search strategies were developed by a medical librarian, HW, in consultation with LY. Strategies combined the general concepts of IBD and AI and endoscopy, using a combination of subject headings and text words relevant to each database. Results were limited to English language and January 1, 2000 onward. Animal studies were excluded. A full electronic search strategy is provided in **Supplementary Table 1**. All searches were run on April 28, 2021.

Study selection

Search results were exported to Endnote X9 (Clarivate, Philadelphia, Pennsylvania, United States). Duplicate records and excluded publication types (book chapters, case reports, comments, conference abstracts and papers, letters and notes) were removed. Remaining records were uploaded to Covidence systematic review software (Veritas Health Innovation, Melbourne Australia) for screening. Two reviewers (LY and EP) independently screened titles and abstracts followed by full text.

Data collection process

Data from included studies were independently collated into a pre-specified data extraction sheet by LY and EP, then cross-checked. Data extracted included year and place of publication, study design, clinical application, size of the training and validation data sets, type of AI algorithm, outcome evaluation metrics and diagnostic performance results.

Risk of bias and applicability

Assessment of risk of bias and applicability was performed by two independent authors (LY and LS) using the Quality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2) [18]. Any discrepancies in the assessment were resolved by consensus.

Diagnostic test accuracy measures

The primary measures of DTA were sensitivity, specificity, positive predictive value, negative predictive value (NPV), accuracy

or Area Under the Receiver Operating Characteristics curve, where data was available.

Synthesis of results

Narrative synthesis was performed for clinical applications. Studies were grouped by IBD disease subtype (ulcerative colitis and CD) and ordered by risk of bias scores. In the absence of an overall rating for the QUADS-2 tool, an overall rank was given for each study based on the four domains. Where the rank was the same, studies were ordered alphabetically using author surname.

A meta-analysis was not performed due to the variations in outcome metrics used and the lack of raw data required for performing the bivariate model or hierarchical summary receiver operating characteristic model for DTA studies [19]. In addition, a number of studies reported individual sensitivity and specificity for subgroup analysis only, without reporting the overall values for the AI model or raw data to calculate true- and false-positive and negative values.

Quantitative synthesis was performed for DTA using median and interquartile range of the reported sensitivity and specificity values. These were charted in box-and-whisker plots. Results from the included studies are presented in summary tables.

Results

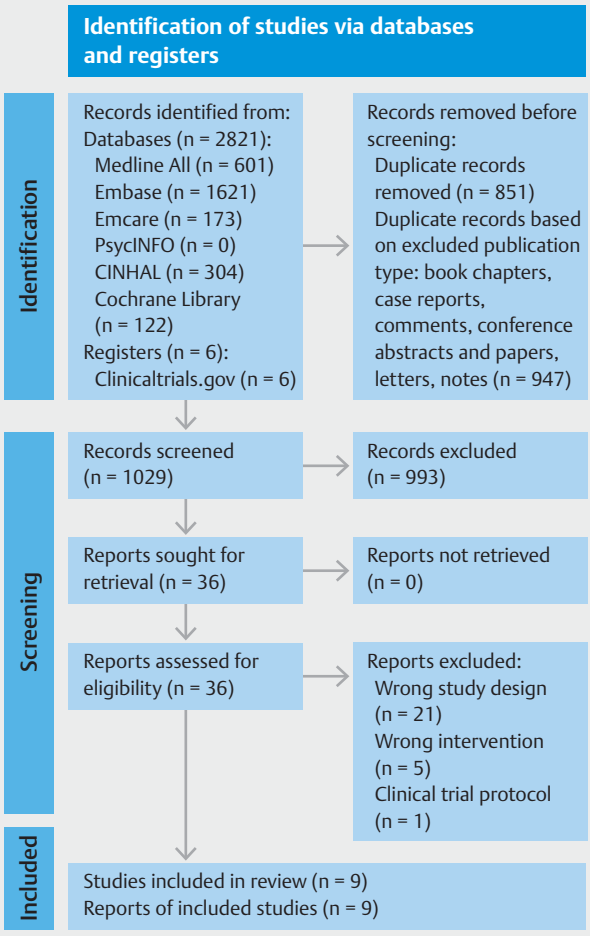
Study selection and characteristics

Of 2821 records identified, nine studies with 7813 patients were included (▶ **Fig. 1**). ▶ **Table 1** summarizes the study characteristics of all articles included in the systematic review, including the study design, the clinical application, size of the dataset and AI algorithm used. Eight studies [20–26] with 7086 patients were on the use of AI in ulcerative colitis (UC) and one study [27] with 727 patients evaluated the use of AI in differentiation of CD from Bechet's disease and intestinal tuberculosis. The first article on the topic of AI in colonoscopic imaging of IBD appeared in 2019. Five studies were conducted in the United States, [20, 23–25, 28] three in Japan, [21, 22, 26] and one in South Korea [27]. Five studies were single-center and retrospective in study design. There were two multicenter retrospective studies. Two studies were prospective and performed in a single center.

All studies used white light endoscopy images. There were no studies using other imaging modalities such as narrow band imaging or chromoendoscopy. There were no studies investigating the use of AI in dysplasia assessment in IBD.

Risk of bias and applicability

According to the QUADAS-2 tool, one article [27] presented high risks of bias, mainly due to patient selection bias and reference standard bias. Most studies used a single-center or publicly available retrospective database or an external data set from clinical trials. Studies of high patient selection bias [25, 27] failed to describe the study population in detail, including the inclusion and exclusion criteria. All articles used an independent validation set of images, from internal [21, 22, 26–28], internal-external [24] or external [20, 25] datasets (▶ **Table 1**).



▶ **Fig. 1** PRISMA flow diagram. Template from: *Page MJ, McKenzie JE, Bossuyt PM* et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021; 372:n71

The risk of bias and applicability assessment outcomes for the included studies are shown in ▶ **Table 2** and **Supplementary Fig. 1**.

Results of individual studies

Study characteristics for individual studies are provided in ▶ **Table 1**. Key findings of individual studies are presented in ▶ **Table 3** and ▶ **Table 4**, including the reference standard, diagnostic performance metrics and estimates of DTA and confidence intervals, where available.

Synthesis: AI algorithms and clinical application

All but one study used a deep neural network (DNN) as the backbone. The most commonly utilized DNN model was the convolutional neural network (CNN). All studies used a retrospective data set of still images or video frames for training of the AI algorithm. All studies included a separate validation or hold-out test set for evaluation of the algorithm. There was significant variation in the types of patient cohort from which the training and validation images were extracted. All still images

▶ **Table 1** Study characteristics.

Author year	Study design	Disease	AI algorithm	Application	Training set images/videos	Reference standard	Validation patients	Validation images/videos	Video
Bhambhvani [20] 2021	Single center, retrospective	UC	CNN (101 layer)	Grading MES	Hyper-Kvasir publicly available retrospective dataset 582	2 endoscopists	NA	116 validation set, 79 hold-out test set	No
Gottlieb [21] 2021	Multi center, retrospective	UC	Bidirectional RNN (5-fold cross validation)	Grading MES and UCEIS, per colon section	629 videos from Phase 2 trial (5.9 million frames)	1 endoscopist	157 Internal-external	157 videos (1.5 million frames)	Yes
Gutierrez Becker [22] 2021	Multi center, retrospective	UC	CNN (5-fold cross validation)	Grading MES per colon section	351 sigmoidoscopy videos (4371 frames) from Phase 2 multicenter RCT	Central reader(s) for clinical trials	1105 External from Phase 3 multicenter RCT	1672 sigmoidoscopy videos (no. of frames not reported) 778 still images from Hyper-Kvasir publicly available retrospective dataset	Yes
Ozawa [23] 2019	Single center, retrospective	UC	CNN (22-layers)	Predicting endoscopic disease activity using MES (0–1 vs 2–3) Correlation between MES and Matts histologic grades	Retrospective database of day procedure clinic 26,304	2–3 endoscopists	114 internal	3981	No
Stidham [25] 2019	Single center, retrospective	UC	CNN (10-fold cross validation)	Predicting endoscopic disease activity using MES (0–1 vs. 2–3)	16 514	2 endoscopists, 3rd reviewer for discrepancies	304 (internal still images) 30 (internal videos)	1652 still images 11 432 images from videos	Yes
Takenaka [19] 2020	Single center, prospective	UC	Deep neural network	Predicting UCEIS and histologic remission	40758	3 endoscopists	875 internal	4187	No
Takenaka [18] 2021	Single center, prospective	UC	Deep neural network	Predicting patient prognosis	40758	3 endoscopists	875 internal	4187	No
Yao [17] 2021	Single center, prospective	UC	CNN (5-fold cross validation)	Grading MES per frame	51 videos (60 frames per second)	2 local endoscopists for training, External central reviewers for validation	157 External from Phase 2 multicenter RCT	264 Videos (no. of frames not reported)	Yes
Kim[24] 2021	Single center, retrospective	Crohn's	CNN	Differentiating Behçet's disease (BD), Crohn's disease (CD), and intestinal tuberculosis (ITB)	5, 237 (2271 BD, 1666 CD, 1300 ITB)	2 endoscopists	697 internal validation set, 683 internal test set	697 validation set (286 BD, 244 CD, 167 ITB) 683 test set (295 BD, 213 CD, 175 ITB)	No

CLE, confocal laser endomicroscopy; EC, endocytoscopyl LSTM, long short-term memory; MES, Mayo endoscopic subscore; NBI, narrow band imaging; RNN, recurrent neural network; UC, ulcerative colitis; UCEIS, UC endoscopic index of severity; WLE, white light endoscopy.

▶ **Table 2** Quality assessment of Diagnostic Accuracy Studies-2 risk of bias assessment.

Study	Patient selection	Index test	Reference standard	Flow and timing	Rank
Bhambhvani et al. [20]	Low	Low	Low	Unclear	2
Gottlieb et al. [21]	Low	Low	High	Unclear	3
Gutierrez Becker et al. [22]	High	Low	Unclear	Low	3
Ozawa et al. [23]	Low	Low	Low	Low	1
Stidham et al. [25]	Low	Low	Low	Low	1
Takenaka et al. [19]	Low	Low	Low	Low	1
Takenaka et al. [18]	Low	Low	Low	Low	1
Yao et al. [17]	Low	Low	Unclear	High	3
Kim et al. [24]	High	Low	High	High	4

▶ **Table 3** Outcomes of artificial intelligence models for prediction of ulcerative colitis using Mayo Endoscopic Subscore.

Author Year	QUADAS-2 Rank	Comparison group	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	PPV (%, 95% CI)	NPV (%, 95% CI)	Accuracy (%)	AUROC (95% CI)	Other
Ozawa et al. [23] 2019	1	Overall	–	–	–	–	MES 0: 73 MES 1: 70 MES 2–3: 63	MES 0 vs 1–3: 0.86 (0.84–0.87) MES 0–1 vs 2–3: 0.98 (0.97–0.98)	–
		With vs without topical treatment	–	–	–	–	–	MES 0 vs 1–3: 0.95 vs 0.91 MES 0–1 vs 2–3: 0.89 vs 0.96	Correlation between Matts grade With topical treatment R = 0.45, *P* = 0.063 Without topical treatment R = 0.42, *P* < 0.0001
		Each location of the colorectum	–	–	–	–	–	MES 0 vs 1–3 Right colon 0.83 Left colon 0.83 Rectum 0.92 MES 0–1 vs 2–3 Right colon 0.99 Left colon 0.99 Rectum 0.94	–
Stidham et al. [25] 2019	1	MES 0–1 vs MES 2–3, images	83 (81–85)	96 (95–97)	86 (85–88)	0.94 (93–95)	MES 0: 89 MES 1: 52 MES 2: 70 MES 3: 74	0.970 (0.967–0.972)	κ 0.84
		MES 0–1 vs MES 2–3, video-based images	–	–	68 (67–69)	0.98 (97–99)	MES 0: 75 MES 1: 68 MES 2: 64 MES 3: 68	0.966 (0.963–0.969)	κ 0.75

▶ **Table 3** (Continuation)

Author Year	QUADAS-2 Rank	Comparison group	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	PPV (%, 95% CI)	NPV (%, 95% CI)	Accuracy (%)	AUROC (95% CI)	Other
Bhambhvani [20] 2021	2	Overall/ Average	72	86	78	87	77	–	–
		MES 1	67	91	74	88	–	0.89	–
		MES 2	86	68	78	80	–	0.86	–
		MES 3	64	97	82	93	–	0.96	–
Gottlieb, 2020 [21]	3	MES 0	88 (82–93)	97 (94–99)	78 (71–84)	98 (96–100)	–	0.92	Endoscopic healing 96% (95% CI, 92–99%) QWK 0.84 (95% CI, 0.79–0.90)
		MES 1	65 (57–72)	92 (88–96)	73 (66–80)	88 (83–93)	–	0.78	
		MES 2	60 (53–68)	77 (70–84)	43 (35–51)	87 (82–92)	–	0.69	
		MES 3	74 (67–81)	95 (92–98)	91 (87–95)	84 (79–90)	–	0.85	
Gutierrez Becker, 2021 [22]	3	AI model trained on raw videos	–	–	–	–	–	Raw videos MCES ≥ 1: 0.84 MCES ≥ 2: 0.85 MCES ≥ 3: 0.85 Still images MCES ≥ 2: 0.82 MCES ≥ 3: 0.83	–
		AI model trained on external dataset still images	–	–	–	–	–	Raw videos MCES ≥ 2: 0.72 MCES ≥ 3: 0.77 Still images MCES ≥ 2: 0.85 MCES ≥ 3: 0.91	–
Yao, 2021 [17]	3	Local validation set with informative image classifier	–	–	–	–	78	–	κ 0.84 (95% CI, 0.75–0.92)
		Local validation set without informative image classifer	–	–	–	–	65	–	κ 0.63 (95% CI, 0.52–0.89)
		External validation set with informative image classifier	MES 0: 50 MES 1: 80 MES 2: 54 MES 3: 67	MES 0: 97 MES 1: 89 MES 2: 78 MES 3: 75	–	–	57	MES 0: 0.95 MES 1: 0.89 MES 2: 0.68 MES 3: 0.71	κ 0.59 (95% CI, 0.46–0.71)
		Segments with a MES of 0 or 1	65 (54–75)	98 (94–99)	87 (76–94)	92 (89–95)	91 (88–94)	–	–
		Per patient assessment	86 (75–94)	93 (80–99)	94 (85–99)	83 (69–92)	89 (81–94)	–	–

AI, artificial intelligence; MES, Mayo endoscopic subscore; κ, Kappa co-efficient; -, not recorded.

▶ **Table 4** Outcomes of artificial intelligence models for prediction of Ulcerative Colitis using Ulcerative Colitis Endoscopic Index of Severity.

Author Year	QUADAS-2 Rank	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	PPV (%, 95% CI)	NPV (%, 95% CI)	Accuracy (%, 95% CI)	AUROC (%, 95% CI)
Takenaka et al. [19] 2020	1	92	91	86	95	–	–
Takenaka et al. [18] 2021	1	Endoscopic remission: 93 (92–94) Histologic remission: 92 (91–93)	Endoscopic remission: 88 (87–88) Histologic remission: 94 (93–94)	Endoscopic remission: 84 (83–85) Histologic remission: 94 (93–95)	Endoscopic remission: 95 (94–96) Histologic remission: 92 (91–93)	Endoscopic remission 90 (89–91) Histologic remission: 93 (92–94)	–
Gottlieb et al. [21] 2021	3	UCEIS 0: 67 UCEIS 1: 60 UCEIS 2: 23 UCEIS 3: 27 UCEIS 4: 33 UCEIS 5: 82 UCEIS 6: 0 UCEIS 7: 0 UCEIS 8: 0	UCEIS 0: 98 UCEIS 1: 86 UCEIS 2: 91 UCEIS 3: 95 UCEIS 4: 84 UCEIS 5: 84 UCEIS 6: 97 UCEIS 7: 100 UCEIS 8: 100	UCEIS 0: 98 UCEIS 1: 92 UCEIS 2: 83 UCEIS 3: 87 UCEIS 4: 93 UCEIS 5: 93 UCEIS 6: 92 UCEIS 7: 99 UCEIS 8: 99	UCEIS 0: 67 UCEIS 1: 43 UCEIS 2: 38 UCEIS 3: 50 UCEIS 4: 17 UCEIS 5: 64 UCEIS 6: 0 UCEIS 7: 0 UCEIS 8: 0	–	UCEIS 0: 0.885 UCEIS 1: 0.333 UCEIS 2: 0.417 UCEIS 3: 0.464 UCEIS 4: 0.492 UCEIS 5: 0.500 UCEIS 6: 0.500 UCEIS 7: 0.500 UCEIS 8: 0.500

AUROC, area under receiver operating characteristics curve; CI, confidence interval; NPV, negative predictive value; PP, positive predictive value; QWK, quadratic weighted kappa; UCEIS, UC endoscopic index of severity; -, not recorded.

for model training were obtained retrospectively from single-center endoscopy databases; all were tertiary institutions except for one study [26] which was a day procedure center. Gutierrez Becker et al. and Gottlieb et al. trained and tested their models using frames from endoscopic videos obtained retrospectively from multicenter clinical trials [24, 25]. Bhambhvani et al. [23] and Gutierrez Becker et al. [24] utilized images from Hyper-Kvasir publicly available retrospective dataset of endoscopy images for training and validation, respectively. Most studies used expert gastroenterologists and/or pathologists as the reference standard for training and validation of their AI algorithms.

Eight studies assessed the use of AI in UC [20–26]. Seven of these studies evaluated computer-assisted prediction of endoscopic and/or histologic disease activity [20, 22–26]. The endoscopic disease assessment tools used in the studies were the MES and the UCEIS. Gottlieb et al. incorporated both scoring systems for predicting endoscopic disease activity [24]. Takenaka et al. used the UCEIS and all other studies used the MES [21, 22]. Earlier studies by Stidham et al. and Ozawa et al. used binary classification to distinguish endoscopic remission from active disease by grouping the MES to 0–1 and 2–3 [26, 28]. Takenaka et al. also classified endoscopic remission, defined as a UCEIS 0, versus any other disease activity (UCEIS ≥ 1) [22]. Subsequent studies graded individual subscores of the MES or UCEIS. Two studies correlated endoscopic disease severity with histology results as reference standard [22, 26].

Takenaka et al. performed a 1-year follow-up study of 875 patients who had comprised the validation cohort of their DNN system, evaluating the association between endoscopic remission predicted by their AI model and patient prognosis [21]. Their results showed that endoscopic mucosal healing predicted by a deep neural network algorithm is associated with lower risks of hospitalization and colectomy.

There was one study on CNN algorithm involving colonoscopic images of CD. Kim et al. evaluated a model for differentiating endoscopic images of colonic CD from other conditions that mimic CD, namely Bechet's disease and intestinal tuberculosis [27]. This study scored high risks of bias in patient selection and reference standard.

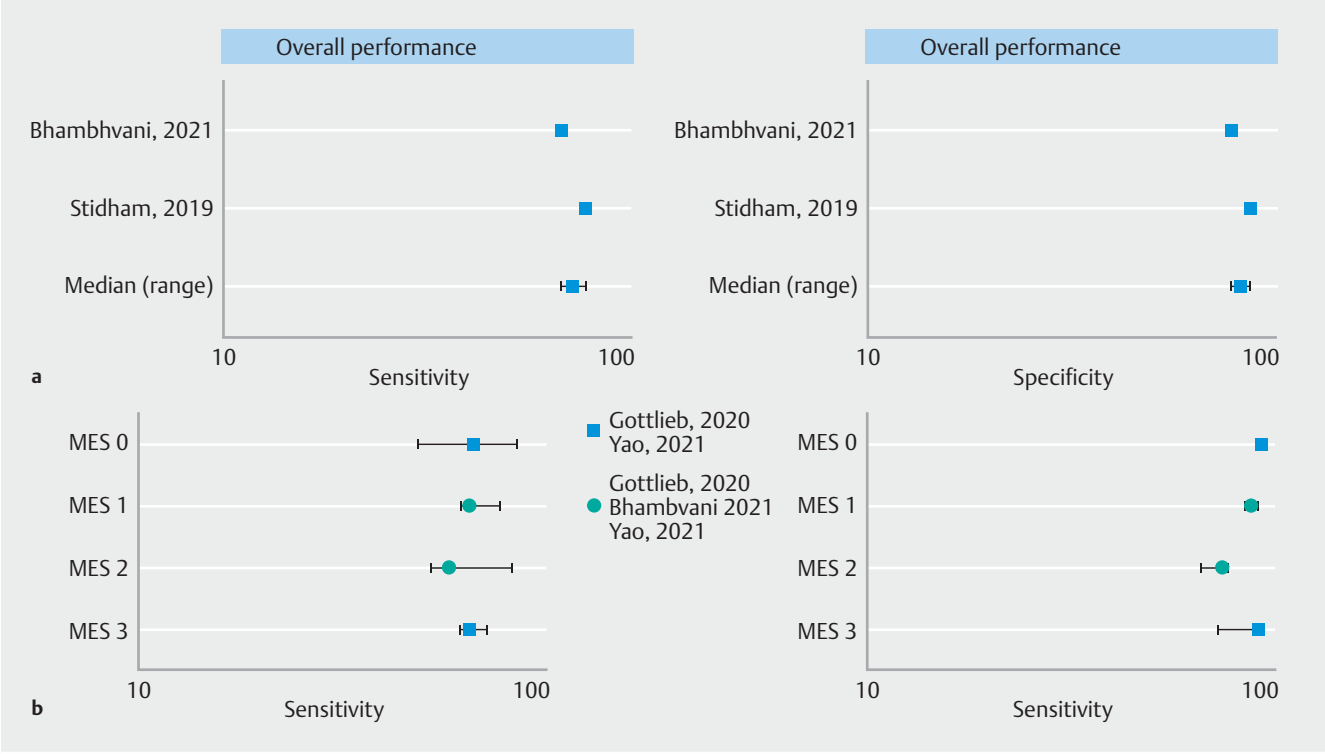
Synthesis: Diagnostic accuracy

AI algorithms for prediction of endoscopic or histologic disease activity in UC performed with an overall sensitivity and specificity of 78% (median, range 72–83, IQR 5.5) and 91% (median, range 86–96, IQR 5), respectively (▶ **Fig. 2**). Sensitivity and specificity for individual subscores of MES and UCEIS showed higher values for disease remission (MES and UCEIS 0) and severe disease (MES 3 and UCES > 5), compared to moderate severity scores. All models performed with excellent AUROC values.

In CD, the CNN by Kim et al. performed with moderate accuracy of 65% and AUROC between 0.785 to 0.859) in differentiating colonoscopy images of the three diseases [27].

Discussion

AI in the form of deep learning is a rapidly evolving field of research. Studies have shown promising results in computer-assisted detection of gastrointestinal pathology and endoscopic image classification. Our review summarizes the current literature on the use of AI-based systems in colonoscopic assessment of IBD. The majority of available data is on prediction of endoscopic disease severity or mucosal healing in UC.



▶ **Fig. 2** Sensitivity and specificity of AI algorithms on ulcerative colitis**. a** Overall results. **b** Average sensitivity and specificity per MES grade.

All of the neural network algorithms reported good DTA, suitable for a potential diagnostic test [20]. However, the diagnostic performance was variable depending on grades of disease severity. The highest DTA was reported for predicting severe disease or complete endoscopic remission. The performance was poor for predicting or differentiating mild to moderate endoscopic severity [24, 26]. As AI algorithms are trained on endoscopic features annotated and demarcated by humans, its output reflects shortcomings of the inputted dataset. The interobserver agreement among gastroenterologists for mild to moderate endoscopic severity has been reported to be poor [5]. Current AI models have been predominantly trained on human reference standards without histologic correlation and therefore reflect similar limitations to the gastroenterologists.

This review identified that risk of bias was high for reference standards defining the "true" disease activity. Most studies graded MES and UCEIS based on image assessment by internal gastroenterologists and only included images where a consensus score was reached. This may weaken the complexity of the algorithm by excluding images that may train for subtle distinctions between grades of disease severity. Studies using central reading of clinical trial videos were limited by the use of average scores assigned to the entire length of the video rather than individual frames [24, 25]. Images from patients with focal severe disease such as distal colitis or proctitis may be inaccurately graded because the severe disease makes up a relatively small fraction of the overall video.

Current evidence is limited by retrospective data that provided images for training and validation of the AI algorithms. Characteristics of inputted images are a significant factor in determining the diagnostic performance of the AI algorithm. The severity of disease would differ greatly between a cohort enrolled in clinical trials of investigational drugs and elective patients presenting to a day procedure center. CNN model trained on single-center endoscopy database performed with higher accuracy when validated on the internal cohort compared to an external cohort from a phase 2 trial of an investigational oral therapy [20]. This reflects the limitation of a single-center retrospective training dataset in achieving wide applicability. Training of the AI model in a prospective, multicenter setting using images from a wide spectrum of disease activity and patients would be ideal in developing an accurate algorithm reflective of real-life practice.

For these deep learning models to be translated into real-time use, the first important step is the application of the model on real-time colonoscopic videos. The main challenge in applying AI to raw full-motion videos is identifying clinically informative frames. Differentiating disease activity from debris, non-specific inflammation and interventional tissue damage from biopsies is routine practice subconsciously performed by the human endoscopist. However, this can be challenging for computational interpretation. Informative image classifier as demonstrated by Yao and colleagues may be useful in this regard [22, 26, 28]. Improvement in speed would also be critical. Twenty-five to 30 frames a second are already achieved in AI models for colonic polyps [12].

Current data suggest that the clinical applicability of computer-aided IBD colonoscopy is limited, despite potential ad-

vantages for its use in clinical trial settings. In clinical trial settings, AI models can help reduce the time and interobserver variability of central reading which requires time-intensive training and video reviews. AI technology may also be beneficial in standardizing the review processes for novel therapies. In routine practice, however, the clinical significance and feasibility of predictive models of disease severity are uncertain. The clinical benefit may be higher in primary or secondary care settings, for the general endoscopists with limited expertise in endoscopic assessment of IBD. Utilizing AI may improve accuracy in discerning mild and moderate disease. However, this is unlikely to lead to significant clinical sequelae.

There is perhaps a greater clinical need for computer-aided models for the screening and detection of IBD-related dysplasia. Colitis-associated dysplasia and neoplasia are often challenging to detect with conflicting data on the efficacy of high-definition white light endoscopy, chromoendoscopy and narrow band imaging (NBI) [4]. Concurrent inflammation, scarring and presence of inflammatory pseudopolyps are some of the challenges in identifying dysplasia. All studies in current literature used white light endoscopy only. Endoscopic assessment of disease severity does not rely heavily on the use of image enhanced endoscopy. However, NBI and dye-based chromoendoscopy are frequently incorporated, particularly for dysplasia surveillance. Future studies on AI models using advanced imaging modalities and AI application in dysplasia detection would potentially lead to significant clinical advantages. However, IBD-related dysplasia presents with a significant variability in lesion characteristics. The prevalence of endoscopically visible lesions is also low and therefore, the development of a robust image dataset for training of AI algorithms may be challenging. This is an area of need for future research in AI application to IBD endoscopy.

Tontini et al. recently published a systematic review on AI in all endoscopic modalities for IBD using narrative synthesis [13]. Studies using novel endoscopic techniques such as confocal laser endomicroscopy [29], endocytoscopy [30], monochromatic light endoscopy [31] and red density system [32] were included. Our systematic review is the first to focus on AI in IBD colonoscopy. In particular, we focused on studies using standard or high-definition white light colonoscopy rather than novel imaging techniques. This was to determine clinical applicability and generalisability of AI in real-life settings. A notable advantage of this review is that a quantitative synthesis was performed, which provides objective interpretation of the literature. Furthermore, we prospectively registered the protocol in PROSPERO (International Prospective Register of Systematic Reviews) and accounted for risk of bias and applicability using the QUADAS-2 tool.

A limitation of this review is that the DTA data were not pooled in a meta-analysis. However, due to heterogeneity in the presented data with some of the studies exhibiting high risks of bias and missing data, meta-analysis was not feasible. Therefore, the overall DTA needs to be considered within these limitations. Current evidence is limited by non-randomized and observational studies, with their potential for confounding and selection biases. However, this review presents a comprehensive and up-to-date summary of the available literature. Limitations of the current AI models and potential strategies for improvement are summarized in ▶ **Table 5**.

Conclusions

In conclusion, there has been a growing interest in the use of AI in IBD endoscopy with most studies in current literature using DNN to predict endoscopic disease severity in UC. The available data supports that AI models may be a promising adjunct to IBD endoscopy, particularly in prediction of disease severity. This suggests that AI has significant potential for clinical application in other critical components of IBD endoscopy such as dysplasia

▶ **Table 5** Summary of limitations of current evidence and potential strategies for improvement.

Limitations	Suggested solutions	Future outcomes
High interobserver variability for differentiating mild and moderate disease	Prospective and multicenter image dataset from various clinical settings (clinical trials, tertiary center, day procedures, primary care) Inclusion of patients on all types of IBD treatments (topical, oral 5ASA, DMARDs and biologics) Correlation of endoscopic disease activity with histologic examination	Wider clinical applicability of AI algorithms
Algorithms are trained on images without imperfections such as debris, tissue damage from biopsies, poor focus	Prospective and multicenter image dataset with higher variation in the types of training images Use of frames from raw full-motion videos	Application of informative image classifier to raw full-motion videos Improved real-life clinical applicability
NBI or dye-based chromoendoscopy images not included	Prospective, multicenter study design Use of NBI or dye-based chromoendoscopy images in the training, testing and validation Broaden inclusion criteria to procedures for indications other than disease assessment (e.g surveillance)	AI models on detection of IBD-related dysplasia

AI, artificial intelligence; DMARDs, disease modifying anti-rheumatic drugs; IBD, inflammatory bowel disease; NBI, narrow band imaging; 5-ASA, 5-aminosalicylic acid.

detection. Further studies in prospective and multicenter settings with more diverse datasets are necessary to simulate real-world practice and for AI to be routinely implemented in clinical endoscopy of IBD.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] GBD 2017 Inflammatory Bowel Diseaes Collaborators. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol 2020; 5: 17–30

[2] Riddell RH, Goldman H, Ransohoff DF et al. Dysplasia in inflammatory bowel disease: standardized classification with provisional clinical applications. Hum Pathol 1983; 14: 931–968

[3] Australia Cancer Council. Surveillance interval for IBD patients. In: Cancer Council Australia. 2019

[4] Laine L, Kaltenbach T, Barkun A et al. SCENIC international consensus statement on surveillance and management of dysplasia in inflammatory bowel disease. Gastroenterology 2015; 148: 639–651 e628

[5] Travis SP, Schnell D, Krzeski P et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut 2012; 61: 535–542

[6] Mohammed Vashist N, Samaan M, Mosli MH et al. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Syst Rev 2018; 1: CD011450

[7] 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: 505–512

[8] Leoncini G, Donato F, Reggiani-Bonetti L et al. Diagnostic interobserver variability in Crohn's disease- and ulcerative colitis-associated dysplasia: a multicenter digital survey from the IG-IBD Pathologists Group. Tech Coloproctol 2021; 25: 101–108

[9] Misawa M, Kudo SE, Mori Y et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018; 154: 2027–2029 e2023

[10] Kudo SE, Misawa M, Mori Y et al. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. Clin Gastroenterol Hepatol 2020; 18: 1874–1881 e1872

[11] Repici A, Badalamenti M, Maselli R et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020; 159: 512–520 e517

[12] Barua I, Vinsard DG, Jodal HC et al. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. Endoscopy 2021; 53: 277–284

[13] Tontini GE, Rimondi A, Vernero M et al. Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons. Therap Adv Gastroenterol 2021; 14: doi:10.1177/17562848211017730

[14] Gubatan J, Levitte S, Patel A et al. Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. World J Gastroenterol 2021; 27: 1920–1935

[15] Sundaram S, Choden T, Mattar MC et al. Artificial intelligence in inflammatory bowel disease endoscopy: current landscape and the road ahead. Ther Adv Gastrointest Endosc 2021; 14: doi:10.1177/26317745211017809

[16] van der Laan JJH, van der Waaij AM, Gabriels RY et al. Endoscopic imaging in inflammatory bowel disease: current developments and emerging strategies. Expert Rev Gastroenterol Hepatol 2021; 15: 115–126

[17] McInnes MDF, Moher D, Thombs BD et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA Statement. JAMA 2018; 319: 388–396

[18] Whiting PF, Rutjes AW, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155: 529–536

[19] Leeflang MM. Systematic reviews and meta-analyses of diagnostic test accuracy. Clin Microbiol Infect 2014; 20: 105–113

[20] Yao H, Najarian K, Gryak J et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc 2021; 93: 728–736 e721

[21] Takenaka K, Ohtsuka K, Fujii T et al. Deep neural network accurately predicts prognosis of ulcerative colitis using endoscopic images. Gastroenterology 2021; 160: 2175–2177 e2173

[22] Takenaka K, Ohtsuka K, Fujii T et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020; 158: 2150–2157

[23] Bhambhvani HP, Zamora A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. Eur J Gastroenterol Hepatol 2021; 33: 645–649

[24] Gottlieb K, Requa J, Karnes W et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology 2021; 160: 710–719 e712

[25] Gutierrez Becker B, Arcadu F, Thalhammer A et al. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther Adv Gastrointest Endosc 2021; 14: doi:10.1177/2631774521990623

[26] Ozawa T, Ishihara S, Fujishiro M et al. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest Endosc 2019; 89: 416–421 e411

[27] Kim JM, Kang JG, Kim S et al. Deep-learning system for real-time differentiation between Crohn's disease, intestinal Behcet's disease, and intestinal tuberculosis. J Gastroenterol Hepatol 2021; 36: 2141–2148

[28] Stidham RW, Liu W, Bishu S et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Netw Open 2019; 2: e193963

[29] Queneherve L, David G, Bourreille A et al. Quantitative assessment of mucosal architecture using computer-based analysis of confocal laser endomicroscopy in inflammatory bowel diseases. Gastrointest Endosc 2019; 89: 626–636

[30] Maeda Y, Kudo SE, Mori Y et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointest Endosc 2019; 89: 408–415

[31] Bossuyt P, De Hertogh G, Eelbode T et al. Computer-aided diagnosis with monochromatic light endoscopy for scoring histologic remission in ulcerative colitis. Gastroenterology 2021; 160: 23–25

[32] Bossuyt P, Nakase H, Vermeire S et al. Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density. Gut 2020; 69: 1778–1786