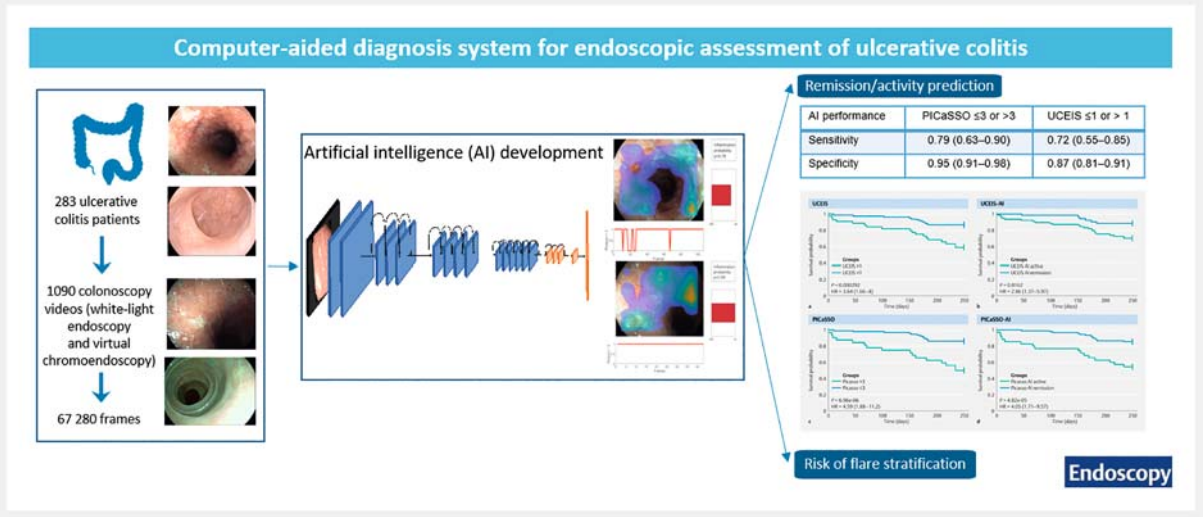
Accepted Manuscript online: 2022-10-13 Article published online: 2022-12-08

A virtual chromoendoscopy artificial intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in ulcerative colitis

OPEN ACCESS

GRAPHICAL ABSTRACT



Authors

Marietta Iacucci[1,2,3], Rosanna Cannatelli[*,1,4], Tommaso L. Parigi[*,1,5], Olga M. Nardone[1,6], Gian Eugenio Tontini[7,8], Nunzia Labarile[9], Andrea Buda[10], Alessandro Rimondi[8], Alina Bazarova[1,11], Raf Bisschops[12], Rocio del Amor[13], Pablo Meseguer[13], Valery Naranjo[13], Subrata Ghosh[1,2,3,14], Enrico Grisan[15,16], on behalf of the PICaSSO group

PICaSSO group

Pradeep Bhandari[17], Gert de Hertogh[12], Jose G. Ferraz[3], Martin Goetz[18], Xianyong Gui[19], Bu'Hussain Hayee[20], Ralf Kiesslich[21], Chiara Metelli[22], Mark Lazarev[23], Remo Panaccione[3], Adolfo Parra-Blanco[24], Luca Pastorelli[25], Timo Rath[26], Elin Synnøve Røyset[27,28], Michael Vieth[29], Vincenzo Villanacci[22], Davide Zardo[30]

Institutions

1 Institute of Immunology and Immunotherapy, NIHR Wellcome Trust Clinical Research Facilities, University of Birmingham, and University Hospitals Birmingham NHS Trust, Birmingham, UK
2 National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK
3 Division of Gastroenterology and Hepatology, University of Calgary, Calgary, Canada
4 Gastroenterology and Digestive Endoscopy Unit, Department of Biochemical and Clinical Sciences "L. Sacco", University of Milan, ASST Fatebenefratelli Sacco, Milan, Italy
5 Department of Biomedical Science, Humanitas University, Milan, Italy
6 Gastroenterology, department of Public health, university of Naples Federico II, Naples, Italy
7 Division of Gastroenterology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
8 Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
9 National Institute of Gastroenterology, IRCCS S. De Bellis Research Hospital, Castellana Grotte, Italy
10 Department of Gastrointestinal Oncological Surgery, Santa Maria del Prato Hospital, Feltre, Italy
11 Institute for Biological Physics, University of Cologne, Cologne, Germany

* Contributed equally to the manuscript.

12 Division of Gastroenterology, University Hospitals Leuven, Leuven, Belgium
13 Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, València, Spain
14 APC Microbiome Ireland, College of Medicine and Health, Cork, Ireland
15 School of Engineering Computer Science and Informatics, London South Bank University, London, UK,
16 Department of Engineering, University of Padova, Padova, Italy
17 Division of Gastroenterology, Queen Alexandra Hospital, Portsmouth, UK
18 Division of Gastroenterology, Klinikum, Böblingen, Germany
19 Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington, USA
20 Division of Gastroenterology, Kings College London, London, UK
21 Helios HSK Wiesbaden, Wiesbaden, Germany
22 Institute of Pathology, Spedali Civili, Brescia, Italy
23 Division of Gastroenterology, Johns Hopkins Hospital, Baltimore, Maryland, USA
24 Division of Gastroenterology, University of Nottingham, Nottingham, UK
25 Liver and Gastroenterology Unit, Department of Health Sciences, Universita' degli Studi di Milano, ASST Santi Paolo E Carlo, University Hospital San Paolo, Milan, Italy
26 Division of Gastroenterology, University of Erlangen, Erlangen, Germany
27 Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Science, Norwegian University of Science and Technology, Trondheim, Norway
28 Department of Pathology at Clinic of Laboratory Medicine, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway
29 Institute of Pathology, Friedrich-Alexander-University Erlangen-Nuremberg, Klinikum Bayreuth, Bayreuth, Germany
30 Department of Pathology, San Bortolo Hospital, Vicenza, Italy

submitted 25.5.2022
accepted after revision 24.8.2022
published online 13.10.2022

Bibliography
Endoscopy 2023; 55: 332–341
DOI 10.1055/a-1960-3645
ISSN 0013-726X

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

Supplementary material
Supplementary material is available under
https://doi.org/10.1055/a-1960-3645

Corresponding author
Marietta Iacucci MD, PhD, Institute of Immunology and Immunotherapy, Heritage Building for Research and Development, University Hospitals Birmingham NHS Foundation Trust, Edgbaston, Birmingham, B15 2TT, UK
m.iacucci@bham.ac.uk

ABSTRACT

Background Endoscopic and histological remission (ER, HR) are therapeutic targets in ulcerative colitis (UC). Virtual chromoendoscopy (VCE) improves endoscopic assessment and the prediction of histology; however, interobserver variability limits standardized endoscopic assessment. We aimed to develop an artificial intelligence (AI) tool to distinguish ER/activity, and predict histology and risk of flare from white-light endoscopy (WLE) and VCE videos.

Methods 1090 endoscopic videos (67 280 frames) from 283 patients were used to develop a convolutional neural network (CNN). UC endoscopic activity was graded by experts using the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and Paddington International virtual ChromoendoScopy ScOre (PICaSSO). The CNN was trained to distinguish ER/activity on endoscopy videos, and retrained to predict HR/activity, defined according to multiple indices, and predict outcome; CNN and human agreement was measured.

Results The AI system detected ER (UCEIS ≤ 1) in WLE videos with 72 % sensitivity, 87 % specificity, and an area under the receiver operating characteristic curve (AUROC) of 0.85; for detection of ER in VCE videos (PICaSSO ≤ 3), the sensitivity was 79 %, specificity 95 %, and the AUROC 0.94. The prediction of HR was similar between WLE and VCE videos (accuracies ranging from 80 % to 85 %). The model's stratification of risk of flare was similar to that of physician-assessed endoscopy scores.

Conclusions Our system accurately distinguished ER/activity and predicted HR and clinical outcome from colonoscopy videos. This is the first computer model developed to detect inflammation/healing on VCE using the PICaSSO and the first computer tool to provide endoscopic, histologic, and clinical assessment.

Introduction

Ulcerative colitis (UC) is a chronic immune-mediated disease characterized by episodes of activity and remission [1]. Over the past decade, there has been an evolution in the treatment targets for UC, from clinical to more objective outcome measures. The first STRIDE consensus [2] established the importance of endoscopic remission (ER) for the maintenance of long-term clinical remission, and the updated STRIDE II [2] introduced the concept of histological remission (HR) as a useful adjunctive measure. The evidence supporting these recommendations arises from a consistent association between deeper mucosal healing and improved clinical outcomes. In contrast, the persistence of inflammatory activity, even when limited to the histological assessment, is associated with increases in flares, hospitalization and, long term, the development of dysplasia [3].

Several definitions of ER have been proposed based on different endoscopic scores. The Mayo Endoscopic Subscore (MES), the first to be introduced, defined ER as a MES ≤ 1 [4]. Since then, other scores such as the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) have been developed and validated to improve the reliability and reproducibility [5]; however, discrepancy persists between ER and HR, largely owing to minimal inflammatory activity being misclassified [3, 6]. Therefore, in clinical practice, biopsies to assess disease activity remain important.

The Paddington International virtual ChromoendoScopy ScOre (PICaSSO) was developed and validated to assess UC mucosal activity and healing with virtual chromoendoscopy (VCE) [7, 8]. VCE enhances mucosal and vascular changes, allowing more accurate characterization of subtle disease activity. Consistent with this, a large multicenter study demonstrated that compared with the MES and UCEIS scores, PICaSSO was more strongly correlated with histological activity and was more accurate in predicting clinical outcomes [9]. Therefore, the advent of VCE has overcome the limitations of WLE, bringing the assessment of endoscopic activity closer to that of histological activity [10].

The major limitation of endoscopic scores is their high inter-rater variability because of the unavoidable subjectivity of the assessments, in spite of improvements in the standardization of training [11]. This is particularly relevant in the context of clinical trials, where central reading has become a necessary countermeasure [12, 13]. To help standardize endoscopic assessment, Takenaka et al. developed a convoluted neural network (CNN) based on an artificial intelligence (AI) system that predicted the degree of inflammation according to the UCEIS; this system was shown to be extremely accurate in replicating endoscopist judgment and predicting histological activity [14–16].

Taking advantage of the accurate prediction of histological activity by VCE and PICaSSO, we aimed to develop an AI-VCE system that was able in real-time to assess ER, and predict HR and increased risk of disease flare on live colonoscopy videos.

Methods

Patients

Patients were recruited from 11 international centers between September 2016 and November 2019 [9]. The inclusion criteria were an established diagnosis of UC for more than 1 year and an indication for endoscopic assessment, regardless of disease activity. The study was approved by the research ethics committee (17/WM/0223) for the centers in the UK, and the local competent committees for the remaining centers.

Endoscopy and videos

All procedures were performed with high definition WLE (HD-WLE) and VCE iSCAN (7010 processor and HiLine series colonoscopes; Pentax, Tokyo, Japan). The colonic mucosa was assessed in HD-WLE and in VCE (iSCAN1, iSCAN2, and iSCAN3). For each patient, two videos with a length of 60–90 seconds each were recorded in the areas of most inflammation or representative of endoscopic healing of the rectum and the sigmoid.

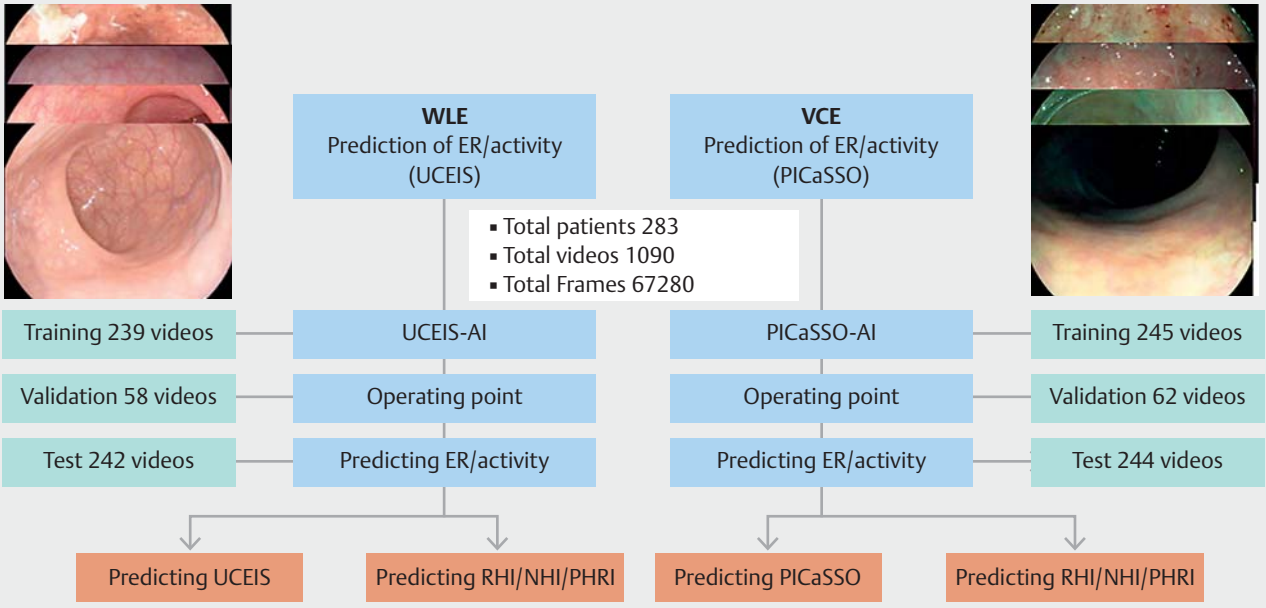
The recordings were edited to separate the sections in WLE and VCE into two different clips, and were annotated and scored by experienced endoscopists from the PICaSSO group of investigators [9]. In HD-WLE videos, endoscopic activity was assessed according to the UCEIS and ER was defined as an UCEIS ≤ 1 [5], whereas VCE videos were assessed with the PICaSSO and ER was defined as a PICaSSO ≤ 3 [9]. In addition, each video clip was graded as high and low quality (LQ), depending on the visibility and clarity of the relevant endoscopic findings. Finally, the edited videos were divided into three sets for training (n = 484), validation (n = 120), and testing (n = 486) of the WLE and VCE systems to predict ER and HR (▶ **Fig. 1**).

Digital pathology

At least two target biopsies were taken from the same areas where the endoscopic assessment was recorded and graded. Samples were fixed in formalin, stained with hematoxylin and eosin (H&E), digitalized at ×40 (0.25 µm per pixel) using the Aperio Digital Pathology Scanning system (Leica Biosystem, Illinois, USA) and assessed by expert pathologists (D.Z., M.V., V. V., G.D.H., E.S.R., and X.G.) who were blinded to clinical information at each center. The histological activity was graded according to the Robarts Histopathology Index (RHI) [17], Nancy Histological Index (NHI) [18], and the newly developed PICaSSO Histologic Remission Index (PHRI) score [19]. HR was defined as an RHI ≤ 3 without neutrophils in the epithelium or lamina propria, NHI ≤ 1, and PHRI = 0.

Clinical outcomes

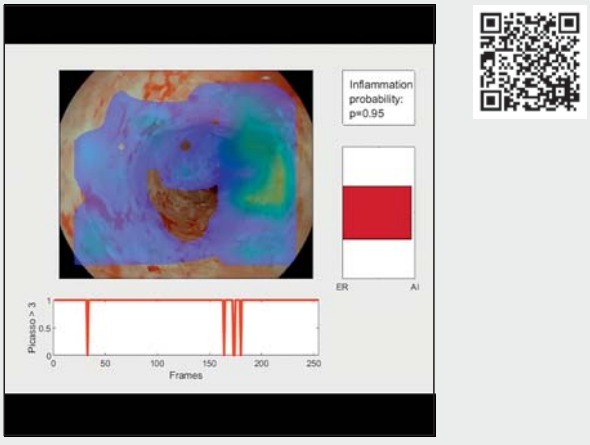
As a proxy of disease flare, data on UC-related hospitalization, colectomy, and initiation or changes in UC therapy (including steroids, immunomodulators, and biological agents) within 12 months after colonoscopy were collected from the clinical records and follow-up phone calls.



▶ **Fig. 1** Development of the artificial intelligence (AI) system involved all endoscopies firstly being edited to separate the white-light endoscopy (WLE) and virtual chromoendoscopy (VCE) parts, then being divided into three sets for training, validation, and testing of the AI models to detect endoscopic remission (ER)/activity according to the UCEIS and PICaSSO, and to predict histological remission, defined by the Nancy Histological Index (NHI), Robarts Histopathology Index (RHI) and PICaSSO Histologic Remission Index (PHRI), and future outcomes.

Artificial intelligence model development

An AI system to analyze endoscopic videos and compose a patient-wide probability of inflammation was developed using HD-WLE and VCE videos clips. The characteristics of the architecture are summarized in ▶ **Fig. 2**. Briefly, the system is based on a transfer learning approach using a ResNet-50 deep residual convolutional neural network (CNN); the network is trained on all frames extracted from videos labelled as containing any signs of endoscopic activity corresponding to a PICaSSO > 3 or UCEIS > 1. When applied on endoscopic videos, the network analyzes each frame as it is acquired, and the frame scores are composed during the video acquisition to provide a patient-wide assessment. To assess histological activity and to predict clinical outcome, the same model was retrained with the same videos associated to new ground truths: histological scores as per pathologist reading, and the occurrence of clinical events as recorded at follow-up (▶ **Fig. 2**; ▶ **Video 1**).



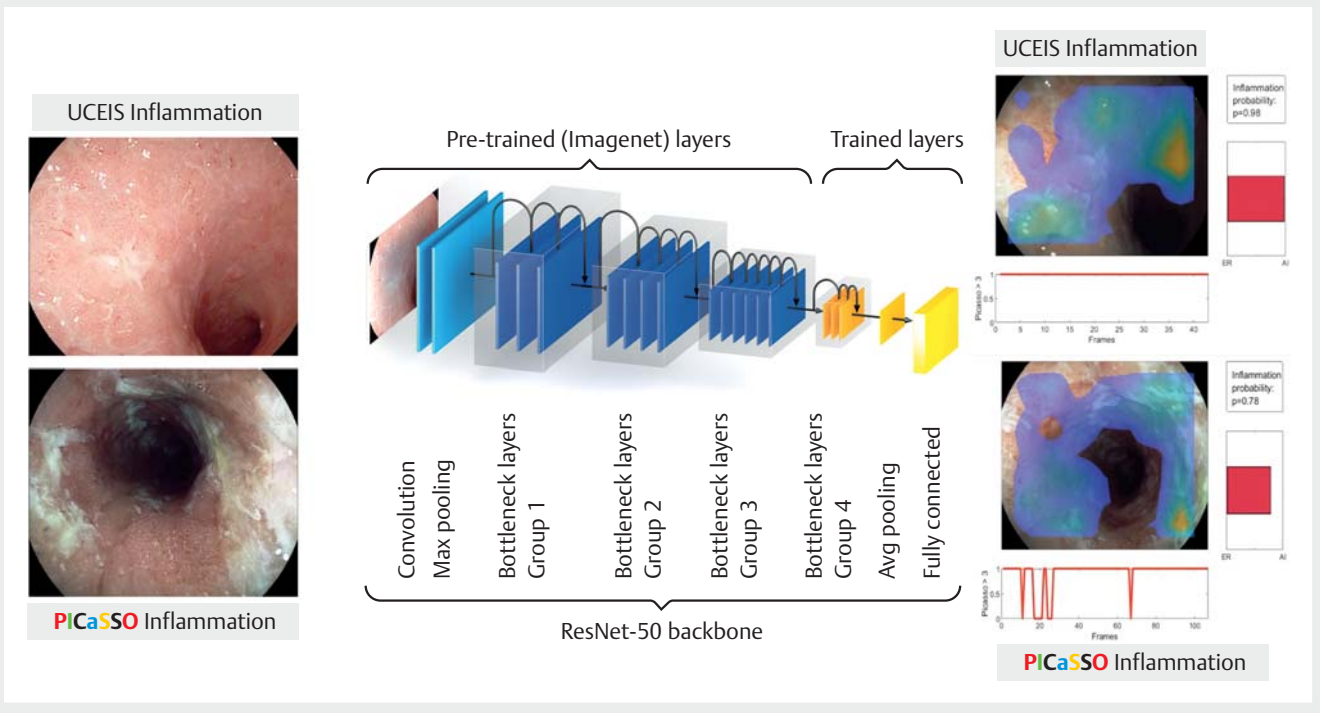
Video 1 Example of the artificial intelligence (AI) system detection of endoscopy remission or activity on high definition white-light endoscopy (HD-WLE) and virtual chromoendoscopy (VCE) videos. All the AI outputs are provided in real time.
Online content viewable at:
https://doi.org/10.1055/a-1960-3645

Objectives

The primary objective of our study was to develop an AI-based computer-aided diagnosis (CAD) system to assess either endoscopic activity or remission. ER was defined as a UCEIS ≤ 1 and PICaSSO ≤ 3 in HD-WLE and VCE videos, respectively.

The secondary objectives were to assess:

1. the ability of the AI CAD system to predict either histological activity or remission; HR was defined as an RHI ≤ 3 without neutrophils in the epithelium and lamina propria, NHI ≤ 1, and PHRI = 0
2. the inter-rater agreement between the CAD system and human endoscopists
3. the ability of the AI CAD system to stratify the risk of incurring prespecified clinical outcomes by 12 months.



▶ **Fig. 2** In the artificial intelligence (AI) architecture, the classification stage of a pretrained ResNet50 convoluted neural network classifier was trained to detect healing or active inflammation on video frames, with two separate networks trained to detect endoscopic remission/activity according to the UCEIS and PICaSSO from frames in high definition white-light endoscopy (HD-WLE) and virtual chromoendoscopy (VCE) videos, respectively. Examples are shown of both HD-WLE and VCE images with features of endoscopic remission and activity that were used to train the model, along with examples of the AI outputs.

Statistical analysis

The sample size was previously calculated for the PICaSSO multicenter study to observe a difference in correlation with histology between PICaSSO and MES [9]. Data were stored in REDCap and analyzed with Matlab (R2021b, The Mathworks Inc., Massachusetts, USA). Continuous variables were reported as mean ± standard deviation (SD). Percentages were calculated and Fisher's exact test or chi-squared statistics were used. The operating point of the AI system (the cutoff value to determine ER/activity) was chosen by means of Youden's J index. To compare humans and AI, contingency tables were prepared and diagnostic performance was reported as sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], accuracy, and area under the receiver operating characteristic curve (AUROC). Confidence intervals were calculated according to Clopper–Pearson [20] for sensitivity, specificity, and accuracy; according to Mercaldo et al. for PPV and NPV [21], and by bootstrapping the data 1000 times and computing the 5th and 95th percentile of the bootstrapped sample for the AUROCs.

The statistical differences in the AUROCs for different classifiers were computed using the nonparametric approach described by DeLong et al. [22]. The agreement among human endoscopic assessments and AI-estimated outputs was measured by Cohen's kappa coefficient: values ≤0 indicating no agreement; 0.01–0.20, none to slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect agreement. Kaplan–Meier survival functions for the two groups of patients (remission versus inflammation) were estimated to evaluate the cumulative risk of incurring any of the specified adverse clinical outcomes (surgery, hospitalization, drug change or optimization) within 12 months. Different survival curves and hazard ratios (HRs) were computed for the groups obtained by the PICaSSO and UCEIS scoring of endoscopists, and the VCE and WLE scoring of the AI system.

The study was conducted and reported following the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) criteria (Table 1 s, see online-only Supplementary material) and the Checklist for Prediction Model Development and Validation (TRIPOD) (Table 2 s).

Results

The demographic characteristics of our study population are summarized in ▶ **Table 1** [9]. Briefly, we included 283 patients, with an average age of 48.2 years (SD 14.8). Around two-thirds of patients were in HR, depending on the biopsy location and histological score used (Table 3 s).

Video collection

Two videos, one in the rectum and one in the sigmoid, were recorded for each of the 283 patients included. After excluding damaged files and recordings where there had been inadequate bowel preparation, the videos were divided into HD-WLE (n = 539) and VCE (n = 551) clips. In total, there were 1090 clips

▶ **Table 1** Demographics and characteristics of the 283 patients included in our study.

Age, mean (SD), years	48.2 (14.8)	
Sex, male, n (%)	165 (58 %)	
Disease duration, mean (SD), years	14.7 (10.0)	
Primary sclerosing cholangitis, n (%)	37 (13 %)	
Extension, n (%)		
▪ Left-sided colitis	122 (43.1 %)	
▪ Subtotal or total colitis	159 (56.2 %)	
▪ Data missing	2 (0.7 %)	
Therapy in last 12 months, n (%)		
▪ No treatment	15 (5.3 %)	
▪ 5-ASA	220 (77.7 %)	
▪ Corticosteroids	71 (25.0 %)	
▪ Immunomodulators	69 (24.4 %)	
▪ Biologics	105 (37.1 %)	
Mayo Endoscopic Score, n (%)		
▪ 0	156 (55.1 %)	
▪ 1	46 (16.3 %)	
▪ 2	52 (18.4 %)	
▪ 3	27 (9.5 %)	
▪ Data missing*	2 (0.7 %)	
UCEIS, n (%)	Rectum	Sigmoid
▪ Remission (≤ 1)	200 (71 %)	208 (73 %)
▪ Active (> 1)	83 (29 %)	75 (27 %)
PICaSSO, n (%)	Rectum	Sigmoid
▪ Remission (≤ 3)	191 (69 %)	221 (78 %)
▪ Active (> 3)	86 (31 %)	62 (22 %)

* Missing data due to inadequate bowel preparation that precluded endoscopic scoring – these patients were not included in the overall analysis.

comprising 67 280 frames, with 901 clips rated as high quality and 189 as low quality. Training, validation, and testing were conducted on a video-wide basis to remove the possible influence of highly correlated frames coming from the same video when reporting the system performance. We assumed that videos from different sections (rectum and sigmoid) of the same patient could be treated as independent.

For HD-WLE, 239 videos were used for the training set, 58 videos for the validation set, and the remaining 242 for testing. For VCE, 245 videos were used for the training set, 62 videos for the validation set, and the remaining 244 for testing. When VCE and HD-WLE videos were available for the same patient and section, they were assigned to the same dataset (training, validation, or testing) for better method comparison. The process is illustrated in ▶ **Fig. 1**.

Primary outcome

Distinguish endoscopic remission (PICaSSO ≤ 3) from activity in VCE

In the testing set, our system detected endoscopic remission/activity (PICaSSO ≤ 3 or > 3) in VCE videos with 79 % (95 %CI 63 %–90 %) sensitivity, 95 % (95 %CI 91 %–98 %) specificity, 77 % (95 %CI 64 %–86 %) PPV, 96 % (95 %CI 92 %–97 %) NPV, 92 % (95 %CI 88 %–95 %) accuracy, and an AUROC of 0.94 (95 %CI 0.91–0.97) (▶ **Table 2**). When restricting the analysis to high quality videos, the sensitivity increased to 86 % (95 %CI 68 %–95 %) and the remaining metrics improved slightly.

Distinguish endoscopic remission (UCEIS ≤ 1) from activity in HD-WLE

For the detection of endoscopic remission/activity in HD-WLE videos (UCEIS ≤ 1 or > 1) in the testing cohort, sensitivity was 72 % (95 %CI 55 %–85 %), specificity 87 % (95 %CI 81 %–91 %), PPV 53 % (95 %CI 43 %–63 %), NPV 94 % (95 %CI 90 %–96 %), accuracy 84 % (95 %CI 79 %–89 %), and AUROC 0.85 (95 %CI 0.79–0.90) (▶ **Table 2**). In the high quality videos subanalysis, sensitivity increased to 79 % (95 %CI 60 %–92 %), specificity to 89 % (95 %CI 83 %–94 %), PPV to 59 % (95 %CI 47 %–70 %), NPV to 96 % (95 %CI 91 %–98 %). The AUROCs of the two AI models, developed on HD-WLE (0.85) and VCE (0.94) videos, were compared using DeLong's test for uncorrelated ROC curves, resulting in a statistically significant difference between the two ($P = 0.02$).

Secondary outcomes

Prediction of histological remission (RHI ≤ 3; NHI ≤ 1; PHRI = 0) from VCE

Our CAD system, analyzing the same videos from VCE, was able to predict HR, defined according to RHI, NHI, and PHRI with accuracies of 83 % (95 %CI 78 %–88 %), 81 % (95 %CI 75 %–86 %), and 83 % (95 %CI 78 %–88 %), respectively, depending on the score used, and AUROCs of 0.83 (95 %CI 0.75–0.90), 0.81 (95 %CI 0.74–0.88), and 0.81 (95 %CI 0.73–0.88) for the same analyses. Regardless of the definition of HR, the accuracy increased by 2 %–3 % when it was restricted to high quality videos only (▶ **Table 3**).

Prediction of histological remission (RHI ≤ 3; NHI ≤ 1; PHRI = 0) from HD-WLE

AI prediction of HR with videos from HD-WLE had accuracies of 80 % (95 %CI 74 %–85 %), 81 % (95 %CI 75 %–86 %), and 80 % (95 %CI 75 %–86 %), and AUROCs of 0.80 (95 %CI 0.72–0.88), 0.81 (95 %CI 0.73–0.88), and 0.79 (95 %CI 0.72–0.87) for RHI, NHI, and PHRI, respectively. When lower quality videos were removed, the accuracy improved by 4 %–5 % (Table 4 s).

Inter-rater agreement between the AI system and human endoscopists

The inter-rater agreement between the AI system and the human endoscopists in detecting ER/activity, expressed as Cohen's kappa coefficient, was substantial (0.73, 95 %CI 0.61–0.85) in VCE videos and moderate (0.51, 95 %CI 0.36–0.66) in

▶ **Table 2** Diagnostic performance in the prediction of endoscopic healing on virtual chromoendoscopy (VCE) using the PICaSSO (≤ 3 or > 3), and on high definition white-light endoscopy (HD-WLE) using the UCEIS (≤ 1 or > 1).

	VCE			HD-WLE		
	Validation	Testing		Validation	Testing	
	62 videos	244 videos	196 high quality videos	58 videos	222 videos	170 high quality videos
Sensitivity (95 %CI)	0.89 (0.66–0.98)	0.79 (0.63–0.90)	0.86 (0.68–0.96)	0.83 (0.61–0.95)	0.72 (0.55–0.85)	0.79 (0.60–0.92)
Specificity (95 %CI)	0.93 (0.81–0.99)	0.95 (0.91–0.98)	0.95 (0.90–0.98)	0.94 (0.81–0.99)	0.87 (0.81–0.91)	0.89 (0.83–0.94)
PPV (95 %CI)	0.85 (0.65–0.94)	0.77 (0.64–0.86)	0.76 (0.61–0.86)	0.90 (0.71–0.97)	0.53 (0.43–0.63)	0.59 (0.47–0.70)
NPV (95 %CI)	0.95 (0.84–0.99)	0.96 (0.92–0.97)	0.98 (0.94–0.99)	0.89 (0.77–0.95)	0.94 (0.90–0.96)	0.96 (0.91–0.98)
Accuracy (95 %CI)	0.92 (0.82–0.97)	0.92 (0.88–0.95)	0.94 (0.89–0.97)	0.90 (0.79–0.96)	0.84 (0.79–0.89)	0.87 (0.81–0.92)
Cohen's kappa (95 %CI)	0.81 (0.66–0.97)	0.73 (0.61–0.85)	0.77 (0.64–0.90)	0.78 (0.61–0.95)	0.51 (0.36–0.66)	0.60 (0.44–0.76)
AUROC (95 %CI)		0.94 (0.91–0.97)			0.85 (0.79–0.90)	

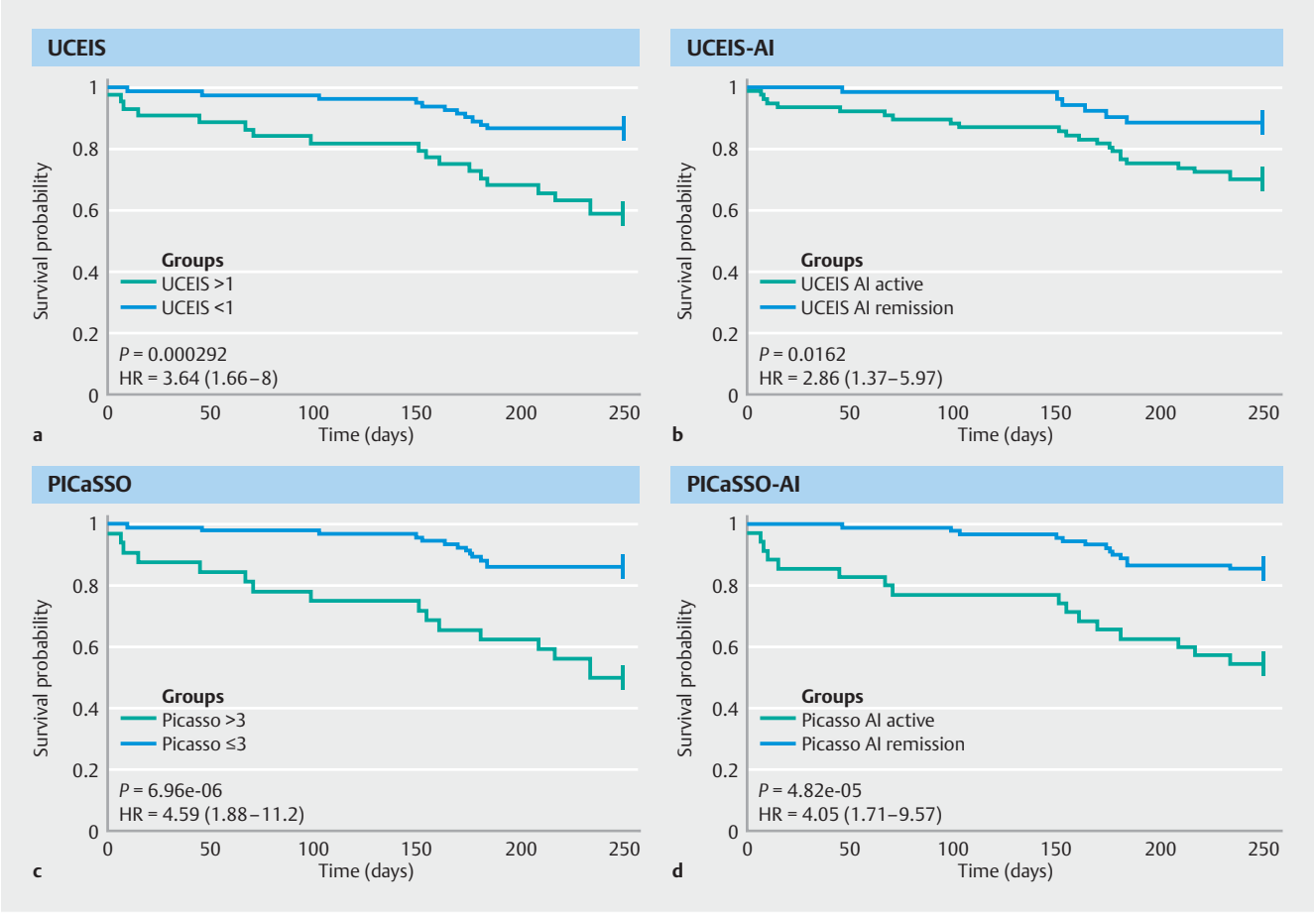
PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristic curve.

▶ **Table 3** Diagnostic performance of the different scores in the prediction of histological healing with virtual chromoendoscopy (VCE) within the testing cohort.

	RHI ≤ 3* or > 3		NHI ≤ 1 or > 1		PHRI ≤ 1 or > 1	
	242 videos	193 high quality videos	242 videos	193 high quality videos	242 videos	193 high quality videos
Sensitivity (95 %CI)	0.73 (0.59–0.85)	0.74 (0.56–0.87)	0.65 (0.51–0.77)	0.64 (0.48–0.78)	0.72 (0.58–0.83)	0.70 (0.54–0.83)
Specificity (95 %CI)	0.86 0.80–0.91)	0.87 (0.81–0.92)	0.86 (0.80–0.91)	0.88 (0.82–0.93)	0.86 (0.81–0.91)	0.88 (0.82–0.93)
PPV (95 %CI)	0.57 (0.47–0.66)	0.57 (0.44–0.66)	0.59 (0.49–0.68)	0.70 (0.48–0.71)	0.62 (0.52–0.71)	0.63 (0.51–0.73)
NPV (95 %CI)	0.93 (0.89–0.95)	0.94 (0.90–0.96)	0.89 (0.85–0.92)	0.90 (0.85–0.93)	0.91 (0.87–0.94)	0.92 (0.87–0.94)
Accuracy (95 %CI)	0.83 (0.78–0.88)	0.85 (0.79–0.90)	0.81 (0.75–0.86)	0.83 (0.77–0.88)	0.83 (0.78–0.88)	0.84 (0.79–0.89)
Cohen's kappa (95 %CI)	0.54 (0.41–0.67)	0.54 (0.39–0.69)	0.49 (0.36–0.62)	0.51 (0.36–0.66)	0.55 (0.43–0.68)	0.55 (0.41–0.70)
AUROC (95 %CI)	0.83 (0.75–0.90)		0.81 (0.74–0.88)		0.81 (0.73–0.88)	

RHI, Robarts Histopathology Index; NHI, Nancy Histological Index; PHRI, PICaSSO Histologic Remission Index; PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristic curve.
* Plus no neutrophils in the lamina propria or epithelium.

HD-WLE videos. Given that the true value of the kappa coefficient lies within the confidence intervals with 95 % probability, agreement for VCE videos is at least substantial, and it is at least fair for HD-WLE videos (▶ **Table 2**). For detection of HR/activity, agreement between the AI CAD and human pathologist was moderate in both sets of videos, VCE and WLE-HQ, ranging between 0.45 and 0.59 (▶ **Table 3**; Table 4 s)



▶ **Fig. 3** Kaplan–Meier survival curves for the two groups of patients (endoscopic remission versus endoscopic activity) to evaluate the cumulative risk of incurring any of the specified adverse clinical outcomes (surgery, hospitalization, drug change or optimization) within 12 months as assessed by the endoscopic scores predicted by: **a,c** human endoscopists; **b,d** the artificial intelligence (AI) model.

AI assessment of risk of prespecified clinical outcomes at 12 months

Of the 283 patients included in the study, 232 patients completed 12 months of follow-up. Of these, 87 suffered one or more of the prespecified adverse clinical outcomes (UC-related hospitalization, colectomy, and UC treatment change owing to relapse). ▶ **Fig. 3** presents the Kaplan–Meier curves for patients in remission or activity according to PICaSSO assessed by human endoscopists (▶ **Fig. 3c**) and the AI system (▶ **Fig. 3d**). For human endoscopists a strong association with risk of outcome for patient with activity is shown (HR 4.59, 95 %CI 1.88–11.2); AI-assessed endoscopic activity was similarly associated with the same outcomes (HR 4.05, 95 %CI 1.71–9.57). The same analysis obtained with HD-WLE classifying remission/activity according to the UCEIS yielded lower hazard ratios (3.64, 95 %CI 1.66–8.0 for human pathologists; 2.86, 95 %CI 1.37–5.97 for AI-assessed endoscopy) (▶ **Fig. 3a,b**).

Bootstrap comparison of the AUROCs for outcome prediction confirmed a statistically significant difference between endoscopist-assessed UCEIS (0.69) and PICaSSO (0.73), and between endoscopist-assessed UCEIS and AI-predicted PICaSSO (0.80). AI-PICaSSO was also numerically superior to AI-UCEIS (0.74), although the difference did not reach statistical significance (Fig. 1 s).

Discussion

The objective and reproducible evaluation of endoscopic activity is crucial to be able to generalize assessment. VCE, through the PICaSSO, has shown the ability to bridge the discrepancy between traditional endoscopic and histological evaluation, allowing the detection of subtle changes overlooked in conventional WLE [23], regardless of the VCE platform [24].

We have developed the first CAD system to evaluate endoscopic and histological activity and remission, and predict specified clinical outcomes through VCE, in addition to conventional HD-WLE, thereby harnessing the potential of image enhancement technology. When applied to VCE videos, our system detected endoscopic inflammatory activity with excellent specificity (95 %) and good sensitivity (79 %). Consistent with the hypothesis that VCE improves optical diagnosis, the same model had slightly worse diagnostic performance with HD-WLE (specificity 87 % and sensitivity 72 %). The statistical comparison of the two AUROCs supports this difference ($P = 0.02$), although caution is necessary because the performances of

the two models (VCE and HD-WLE) are assessed with different scores and cutoffs (PICaSSO ≤3 for VCE and UCEIS ≤1 for HD-WLE). We chose not to use the MES as it is not fully validated, its ER definition includes 0 or 1, and, as several studies have shown, its correlation with histology is lower than that of the PICaSSO and UCEIS [6, 9].

In real-time, our CAD system can provide an initial assessment of inflammation when using HD-WLE and can then support a more accurate evaluation after switching to VCE, which increases the contrast between healthy and inflamed tissue, improving diagnostic performance and requiring only passive confirmation of inflammation or healing by the endoscopists. If the AI-predicted endoscopic activity from VCE were trusted, only 5% (10/202) of remission videos would be misclassified as activity and possibly overtreated. The chance of the opposite error, activity mistaken for remission, would be 21% (9/42 videos from 8 patients), or 14% if considering only high quality videos. Of the eight patients at risk of undertreatment, three suffered a disease flare during follow-up.

In the future, our system could be successfully implemented in both nonexpert and expert clinical practice, as well as in clinical trials. When using the AI model to predict histology, the specificity remained strong (>80%), suggesting that the inflammatory activity seen on endoscopy corresponds to that found in the histology. In contrast, the sensitivities ranged between 66% and 74%, depending on the score, supporting the common notion that some features of histological inflammation are not visible with endoscopy. Overall, however, the diagnostic accuracy in determining HR remained good and greater than 80%.

The similar diagnostic performance of the CAD system in predicting histological activity with VCE and HD-WLE has different possible explanations. First and foremost, VCE improves the detection of inflammation by human endoscopists, but there is no guarantee that an algorithm derives its predictions from the same mucosal features that humans use. Secondly, even if it did, the system might also detect subtle changes in HD-WLE without the need for optical enhancement. The results show that inter-rater agreement between AI and human endoscopists was substantial for VCE and moderate for HD-WLE. Although different scores prevent a direct comparison, the results suggest that assessment using VCE might be more reproducible.

Prediction of prognosis represents an exciting further step in the development of computer tools. The HRs of suffering an adverse clinical outcome in the ER and endoscopic activity groups identified by humans and VCE-AI point to an accurate stratification of the risk of flare. The same classification using HD-WLE/UCEIS was slightly less robust, although caution is necessary as the definitions of endoscopic remission (UCEIS ≤1 and PICaSSO ≤3) are different. Altogether, we expect the accuracy of this type of prediction to increase as larger datasets become available and the system is further refined.

Our work has several strengths. Firstly, to the best of our knowledge, this is the first AI model developed for the assessment of colonoscopy videos based on an optical enhancement system and using several endoscopic and histological scores. The robustness of the dataset is another important factor. Because the PICaSSO study aimed to stress the association between endoscopy and histology, biopsies were matched to the very same areas where the videos were recorded and the endoscopic scores derived. This apparently simple shrewdness is seldom found in other works and reinforces our observations. Furthermore, our cohort of patients was prospectively enrolled, avoiding possible selection or retrieval bias that could have occurred in other studies [14, 25].

Secondly, and important for clinical practice, our AI model is designed to assess whole videos, considered the state-of-the-art approach, rather than single still frames. Although videos are made of frames, the endoscopist's assessment remains based on the entire procedure. To resemble human judgement, we designed our system to detect the most relevant features of the video and ignore frames with milder signs of activity, no signs of disease, or poor image quality, in order to provide a unique result. This approach might sacrifice some diagnostics accuracy, as compared with others, notably the work of Takenaka et al. [14], but it allows a practical use that is more similar to real-life clinical observation, while avoiding the discontinuity and possible selection bias of assessing selected pictures. Moreover, the computerized analysis can take place in real time (see ▶ **Video 1**) or later, providing, on request, a simple and immediately available result to the clinician. Because the video interface shows which areas are identified as inflamed, this ensures the results remain interpretable, a feature often missing in "black box" AI systems.

Thirdly, overfitting is a major concern in AI development. An unsupervised, or loosely supervised, machine-learning model trained with too homogeneous data might underperform when applied to a different setting. This happens because the AI learns from associations that are relevant in a training setting, but may result from what data are presented and how (i.e. if dye is only used in quiescent patients, the algorithm might predict remission from the presence of the dye rather than from the mucosal appearance). This applies also to aspects such as video capture, lighting, and recording. The multicenter source of data (11 centers in 6 countries, each with differences in population and recording equipment) is a major strength and reduces the risk of overfitting.

Our work has some potential limitations. Firstly, all procedures were carried out in tertiary centers by endoscopists experienced in the optical diagnosis of inflammatory bowel disease, which is potentially less representative of ordinary care settings. Secondly, the dataset was limited to the rectum and sigmoid. Nevertheless, given the distribution of UC, the absence of more proximal segments is unlikely to impact the functioning of the model [26]. Videos were of differing quality and this may have affected the diagnostic performance. In fact, unsurprisingly, after removing lower quality clips, the model's performance increased. In addition, the system has not yet been assessed on its responsiveness to treatment. Finally, our model was developed and tested with videos recorded only with the iScan (Pentax) platform. We recently reported that PICaSSO is valid for other optical enhancement platforms [24].

Nevertheless, a prospective multicenter study to validate the system on other VCE platforms is planned.

In conclusion, we developed and tested an AI system to distinguish endoscopic and histological activity from remission in patients with UC using colonoscopy videos from both HD-WLE and VCE. The CAD system developed on VCE videos showed a higher diagnostic performance for the assessment of endoscopic activity compared with the same system based on HD-WLE videos. This tool has multiple potential applications, such as standardizing the assessment of disease activity in daily practice, providing a central readout for clinical trials, supporting less experienced endoscopists, and guiding physicians to target biopsies to the most affected areas. Building on our previous work on computerized assessment of UC histopathology [19], we plan to integrate the two tools and further validate them in a large multicenter study.

Competing Interests

R. Bisschops has received funding, consultancy and speaker's assignments from Pentax, Fujifilm and Medtronic. M. Iacucci is partially funded by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health. A. Bazarova, A. Buda, R. Cannatelli, R. del Amor, S. Ghosh, E. Grisan, N. Labarile, P. Meseguer, V. Naranjo, O.M. Nardone, T.L. Parigi, A. Rimondi, and G.E. Tontini declare that they have no conflict of interest.

References

[1] Ungaro R, Mehandru S, Allen PB et al. Ulcerative colitis. Lancet 2017; 389: 1756–1770

[2] Turner D, Ricciuto A, Lewis A et al. STRIDE-II: an update on the selecting therapeutic targets in inflammatory bowel disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021; 160: 1570–1583

[3] Yoon H, Jangi S, Dulai PS et al. Incremental benefit of achieving endoscopic and histologic remission in patients with ulcerative colitis: a systematic review and meta-analysis. Gastroenterology 2020; 159: 1262–1275.e7

[4] Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. NEJM 1987; 317: 1625–1629

[5] Travis SPL, 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] Bryant RV, Burger DC, Delo J et al. Beyond endoscopic mucosal healing in UC: histological remission better predicts corticosteroid use and hospitalisation over 6 years of follow-up. Gut 2016; 65: 408–414

[7] Iacucci M, Daperno M, Lazarev M et al. Development and reliability of the new endoscopic virtual chromoendoscopy score: the PICaSSO (Paddington International Virtual ChromoendoScopy ScOre) in ulcerative colitis. Gastrointest Endosc 2017; 86: 1118–1127.e5

[8] Trivedi PJ, Kiesslich R, Hodson J et al. The Paddington International Virtual Chromoendoscopy Score in ulcerative colitis exhibits very good inter-rater agreement after computerized module training: a multicenter study across academic and community practice (with video). Gastrointest Endosc 2018; 88: 95–106.e2

[9] Iacucci M, Smith SCL, Bazarova A et al. An international multicenter real-life prospective study of electronic chromoendoscopy score PICaSSO in ulcerative colitis. Gastroenterology 2021; 160: 1558–1569.e8

[10] Nardone OM, Cannatelli R, Zardo D et al. Can advanced endoscopic techniques for assessment of mucosal inflammation and healing approximate histology in inflammatory bowel disease? Therap Adv Gastroenterol 2019; 12: 1756284819863015

[11] Fernandes SR, Pinto JSLD, Marques da Costa P et al. Disagreement among gastroenterologists using the Mayo and Rutgeerts Endoscopic Scores. Inflamm Bowel Dis 2018; 24: 254–260

[12] Gottlieb K, Requa J, Karnes W et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology 2021; 160: 710–719.e2

[13] Gottlieb K, Daperno M, Usiskin K et al. Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments. Gut 2021; 70: 418–426

[14] 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

[15] Takenaka K, Fujii T, Kawamoto A et al. Deep neural network for video colonoscopy of ulcerative colitis: a cross-sectional study. Lancet Gastroenterol Hepatol 2022; 7: 230–237

[16] 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.e3

[17] Mosli MH, Feagan BG, Zou G et al. Development and validation of a histological index for UC. Gut 2017; 66: 50–58

[18] Marchal-Bressenot A, Salleron J, Boulagnon-Rombi C et al. Development and validation of the Nancy histological index for UC. Gut 2017; 66: 43–49

[19] Gui X, Bazarova A, Del Amor R et al. PICaSSO Histologic Remission Index (PHRI) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system. Gut 2022; 71: 889–898

[20] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934; 26: 404–413

[21] Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. Stat Med 2007; 26: 2170–2183

[22] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837–845

[23] Nardone OM, Bazarova A, Bhandari P et al. PICaSSO virtual electronic chromendoscopy accurately reflects combined endoscopic and histological assessment for prediction of clinical outcomes in ulcerative colitis. United European Gastroenterol J 2022; 10: 147–159

[24] Cannatelli R, Bazarova A, Furfaro F et al. Reproducibility of the electronic chromoendoscopy PICaSSO score (Paddington International Virtual ChromoendoScopy ScOre) in ulcerative colitis using multiple endoscopic platforms: A prospective multicenter international study. Gastrointest Endosc 2022; 96: 73–83

[25] 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.e1

[26] Colombel J-F, Ordás I, Ullman T et al. Agreement between rectosigmoidoscopy and colonoscopy analyses of disease activity and healing in patients with ulcerative colitis. Gastroenterology 2016; 150: 389–395.e3