Endoscopy

A Virtual Chromoendoscopy Artificial Intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in Ulcerative Colitis


Affiliations below.

DOI: 10.1055/a-1960-3645

Please cite this article as: Iacucci M, Cannatelli R, Parigi T et al. A Virtual Chromoendoscopy Artificial Intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in Ulcerative Colitis. Endoscopy 2022. doi: 10.1055/a-1960-3645

Conflict of Interest: Marietta 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.

Prof. Raf Bisschops has received funding, consultancy and speaker's assignments from Pentax, Fujifilm and Medtronic. The remaining authors report no relevant conflict of interest relevant to this manuscript.

Abstract:

Background and study aims
Endoscopic and histologic remission (ER, HR) are therapeutic targets in ulcerative colitis (UC) and virtual chromoendoscopy (VCE) improves the endoscopic assessment and the prediction of histology. However, interobserver variability is a limitation for widespread standardised endoscopic assessment using all scoring systems. We aimed to develop an artificial intelligence tool to distinguish ER/activity, and predict histology and outcomes from white-light-endoscopy (WLE) and VCE videos.

Patients and methods
1090 endoscopic videos (638287 frames), from 283 patients, were used to develop a convolutional neural network (CNN). UC endoscopic activity was graded by experts with Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and 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 humans agreement was measured.

Results
The AI system detected ER (UCEIS ≤1) in WLE videos with 72% sensitivity (Se), 87% specificity (Sp), and area under the ROC curve (AUROC) of 0.85; For detection of ER in VCE videos (PICaSSO ≤3) Se was 79%, Sp 95%, and the AUROC 0.94. Prediction of HR was similar between WLE and VCE videos (accuracies ranging 80%-85%). The model's prediction of clinical outcomes was similar to that of physician-assessed endoscopy scores.

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

Corresponding Author:
Marietta Iacucci, University of Birmingham School of Clinical and Experimental Medicine, Institute of Immunology and Immunotherapy, Birmingham, United Kingdom of Great Britain and Northern Ireland, m.iacucci@bham.ac.uk

Affiliations:
Marietta Iacucci, University of Birmingham School of Clinical and Experimental Medicine, Institute of Immunology and Immunotherapy, Birmingham, United Kingdom of Great Britain and Northern Ireland
Marietta Iacucci, NIHR Birmingham Biomedical Research Centre, Division of Gastroenterology, Birmingham, United Kingdom of Great Britain and Northern Ireland
Marietta Iacucci, University of Calgary, Division of Gastroenterology and Hepatology, Calgary, Canada
Enrico Grisan, University of Padova Department Industrial Engineering, Department of Engineering, Padova, Italy
A Virtual Chromoendoscopy Artificial Intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in Ulcerative Colitis

Authors: M. Iacucci 1, 2, 3, R. Cannatelli 1, 4, T.L. Parigi 1, 5, O.M. Nardone 1, 6, G.E. Tontini 7, 8, N. Labarile 9, A. Buda 10, A. Rimondi 8, A. Bazarova 1, 11, R. Bisschops 12, R. del Amor 13, P. Meseguer 13, V. Naranjo 13, PICaSSO Group†, S. Ghosh 1, 2, 3, 14, E. Grisan 15, 16

*Contributed equally to the manuscript

Affiliations:
1Institute of Immunology and Immunotherapy, NIHR Wellcome Trust Clinical Research Facilities, University of Birmingham, and University Hospitals Birmingham NHS Trust, Birmingham, United Kingdom
2National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, United Kingdom,
3Division of Gastroenterology and Hepatology, University of Calgary, Calgary, Canada
4Gastroenterology and Digestive Endoscopy Unit, Department of Biochemical and Clinical Sciences "L. Sacco", University of Milan, ASST Fatebenefratelli Sacco, Milan, Italy,
5Department of Biomedical Science, Humanitas University, Milan, Italy
6Gastroenterology, Department of Clinical Medicine and Surgery, University Federico II of Naples, Naples, Italy
7Division of Gastroenterology, Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milan, Italy
8Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
9Section of Gastroenterology II, National Institute of Research "Saverio De Bellis", Castellana Grotte, Italy
10Department of Gastrointestinal Oncological Surgery, Santa Maria del Prato Hospital, Feltre, Italy
11Institute for Biological Physics, University of Cologne, Cologne, Germany
12Division of Gastroenterology, University Hospitals Leuven, Leuven, Belgium
13Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, València, Spain
14APC Microbiome Ireland, College of Medicine and Health, Cork, Ireland
School of Engineering Computer Science and Informatics, London South Bank University, London, United Kingdom,

Department of Engineering, University of Padova, Padova, Italy


Division of Gastroenterology, Queen Alexandra Hospital, Portsmouth, United Kingdom

Department of Pathology and Laboratory Medicine, University of Calgary, Calgary, Alberta, Canada

Division of Gastroenterology, Klinikum, Böblingen, Germany

Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington, United States

Division of Gastroenterology, Kings College London, London, United Kingdom

Division of Gastroenterology, Johns Hopkins Hospital, Baltimore, Maryland, United States

Division of Gastroenterology, University of Nottingham, Nottingham, United Kingdom

Liver and Gastroenterology Unit, Department of Health Sciences, Universita' degli Studi di Milano, ASST Santi Paolo E Carlo, University Hospital San Paolo, Milan, Italy

Division of Gastroenterology, University of Erlangen, Erlangen, Germany

Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Institute of Pathology, Friedrich-Alexander-University Erlangen-Nuremberg, Klinikum Bayreuth, Bayreuth, Germany

Department of Pathology, San Bortolo Hospital, Vicenza, Italy

Correspondence

Address for correspondence:
Prof. Marietta Iacucci MD, PhD, FASGE, AGAF
Institute of Immunology and Immunotherapy
Heritage Building for Research and Development
University Hospitals Birmingham NHS Foundation Trust
Funding
MI and SG are 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(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests
None of the other authors have any conflict of interest to declare related to this manuscript.

Patient consent for publication
Not required

Ethics approval
The study was approved by the West Midlands Research Ethics Committee (17/WM/0223). All patients gave informed consent to participate in the study.

Abbreviations
IBD, inflammatory bowel disease; UC, ulcerative colitis; CD, Crohn’s disease; WLE, white light endoscopy; HD, high definition; VCE, virtual chromoendoscopy; ER, endoscopic remission; HR, histologic remission; UCEIS, ulcerative colitis endoscopic index of severity; MES, Mayo endoscopic score; PICASSO, Paddington International virtual ChromoendoScopy ScOre; RHI, Robart Histopathology index; NHI, Nancy Histological index; PHRI, Picasso Histologic remission index; CAD, Computer-aided design; AI, Artificial Intelligence; CNN, convolutional neural network;

Authors contributions
MI: Study conception and design, Concept of AI for endoscopy scoring, Data acquisition, Analysis and interpretation of data, Drafting of the manuscript, Critical revision of manuscript for important intellectual content
RC: Data acquisition, Analysis and interpretation of data, Drafting of the manuscript, Critical revision of manuscript for important intellectual content
TLP: Data acquisition, Analysis and interpretation of data, Drafting of the manuscript, Critical revision of manuscript for important intellectual content
M.N: Data acquisition, Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
GET: Data acquisition, Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
N.L: Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
AB: Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
AR: Data acquisition, Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
A.B: Analysis and interpretation of data, Critical revision of manuscript for important intellectual content
P.B: Data acquisition, Critical revision of manuscript for important intellectual content
R.B: Data acquisition, Critical revision of manuscript for important intellectual content
G.DDH: Data acquisition, Critical revision of manuscript for important intellectual content
R.DA: Data acquisition, Critical revision of manuscript for important intellectual content
J.G.F: Data acquisition, Critical revision of manuscript for important intellectual content
M.G: Data acquisition, Critical revision of manuscript for important intellectual content
X.G: Data acquisition, Interpretation of data, Critical revision of manuscript for important intellectual content
B.H: Data acquisition, Critical revision of manuscript for important intellectual content
R.K: Critical revision of manuscript for important intellectual content
M.L: Data acquisition, Interpretation of data, Critical revision of manuscript for important intellectual content
V.N: Interpretation of data, Critical revision of manuscript for important intellectual content
R.P: Interpretation of data, Critical revision of manuscript for important intellectual content
A.P: Data acquisition, Interpretation of data Critical revision of manuscript for important intellectual content
L.P: Data acquisition, Interpretation of data Critical revision of manuscript for important intellectual content
T.R: Data acquisition, Interpretation of data Critical revision of manuscript for important intellectual content
E.S.R: Interpretation of data Critical revision of manuscript for important intellectual content
M.V: Data acquisition Interpretation of data Critical revision of manuscript for important intellectual content
V. V: Interpretation of data Critical revision of manuscript for important intellectual content
D. Z: Data acquisition, Interpretation of data Critical revision of manuscript for important intellectual content
S.G: Study conception and design, Concept of AI for endoscopy scoring, Data acquisition, Analysis and interpretation of data, Drafting of the manuscript, Critical revision of manuscript for important intellectual content
E.G: Study conception and design, Concept of AI for endoscopy scoring, Data acquisition, Analysis and interpretation of data, Drafting of the manuscript, Critical revision of manuscript for important intellectual content
Abstract

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

Patients and methods
1090 endoscopic videos (638287 frames), from 283 patients, were used to develop a convolutional neural network (CNN). UC endoscopic activity was graded by experts with Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and 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 humans agreement was measured.

Results
The AI system detected ER (UCEIS ≤1) in WLE videos with 72% sensitivity (Se), 87% specificity (Sp), and area under the ROC curve (AUROC) of 0.85; For detection of ER in VCE videos (PICaSSO ≤3) Se was 79%, Sp 95%, and the AUROC 0.94. Prediction of HR was similar between WLE and VCE videos (accuracies ranging 80%-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 outcomes from colonoscopy videos. This is the first computer model developed to detect inflammation/healing using VCE through the PICaSSO score and the first computer tool providing endoscopic, histologic, and clinical assessment.
INTRODUCTION

Ulcerative colitis (UC) is a chronic immune-mediated disease characterised by episodes of activity and remission.[1] Over the past decade, there has been an evolution in the treatment targets in 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 (MH) and improved clinical outcomes. On the contrary, the persistence of inflammatory activity, even when limited to the histological assessment, is associated with increases in flares, hospitalisation and, longterm, the development of dysplasia.[3]

Several definitions of ER have been proposed based on different endoscopic scores. Mayo Endoscopic Subscore (MES), the first to be introduced, defined ER as MES ≤1.[4] Afterward, other scores such as Ulcerative Colitis Endoscopic Index of Severity (UCEIS) have been developed and validated to improve the reliability and reproducibility.[5] However, discrepancy persists between endoscopic and histologic remission, largely due to minimal inflammatory activity being misclassified.[3,6] Hence, 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 VirtualChromoEndoscopy (VCE).[7,8] VCE enhances mucosal and vascular changes allowing more accurate characterisation of subtle disease activity. Consistent with this, a large multicentre study demonstrated that compared to MES and UCEIS scores, PICaSSO was more strongly correlated with histologic activity and was more accurate in predicting clinical outcomes.[9] Therefore, the advent of VCE has overcome the limitations of WLE bringing endoscopic activity closer to histologic activity.[10]

The major limitation of endoscopic scores is their high inter-rater variability due to the unavoidable subjectivity of the assessments despite improvements in standardisation 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 standardise 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 UCEIS. This system was
proved to be extremely accurate in replicating the endoscopist judgment[10] and predicting histological activity.[14-16]

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

MATERIALS AND METHODS

Patients
Patients were recruited from 11 international centres between September 2016 and November 2019.[9] 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 UK centres, and local committees.

Endoscopy and Videos
All procedures were performed with WLE-high definition (HD) and VCE iSCAN (7010 processor and HiLine series colonoscopes, Pentax, Tokyo, Japan). The colonic mucosa was assessed in WLE-HD and in VCE (iSCAN1, iSCAN2, and iSCAN3). For each patient two videos, with a length of 60-90 seconds each, were recorded, in the most inflamed areas or representative of endoscopic healing of the rectum and of the sigmoid. The recordings were edited to separate the sections in WLE and VCE into two different clips and annotated and scored by experienced endoscopists from the PICaSSO group of investigators.[9] In HD-WLE videos endoscopic activity was assessed according to UCEIS and ER was defined as UCEIS ≤1.5 whereas VCE videos were assessed with the PICaSSO score and ER was defined as PICaSSO ≤3.[9] In addition, each video clip was graded as high (HQ) and low (LQ) depending on the visibility and clarity of relevant endoscopic findings. Finally, the edited videos were divided into three sets for training (484), validation (120), and testing (486) of the WLE and VCE system to predict ER and HR. (Figure 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 H&E, digitalized at 40× (0.25 μm per pixel) using Aperio Digital Pathology Scanning system (Leica Biosystem, Illinois, USA) and assessed by expert pathologists (DZ, MV, VV, GDH, ESR and XG), blind to clinical
information, at each centre. The histological activity was graded according to Robarts Histological Index (RHI)[17], Nancy Histologic Index (NHI)[18] and the newly developed PHRI score.[19] HR was defined as RHI ≤3 without neutrophils in the epithelium or lamina propria, NHI ≤ 1, and PHRI =0, respectively.

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

Artificial intelligence model development
An AI system to analyse endoscopic videos and compose a patient-wide probability of inflammation was developed using WLE-HD and VCE videos clips. Characteristics of the architecture are detailed in the appendix and summarised in figure 2. Briefly, the system is based on a transfer learning approach using a ResNet-50 deep residual convolutional neural network; the network is trained on all frames extracted from videos labelled as containing any signs of endoscopic activity corresponding to PICaSSO>3 or UCEIS>1. When applied on endoscopic videos, the network analyses 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 occurrence of clinical events as recorded at follow-up. Figure 2 and video

Objectives
The primary objective of our study was to develop an AI-based CAD system to assess endoscopic activity/remission defined as UCEIS ≤1 and PICaSSO ≤3, in WLE-HD and VCE videos, respectively.

The secondary objectives were the following:
- to assess the ability of the AI CAD system to predict histological remission/activity defined as RHI ≤3 without neutrophils in the epithelium and lamina propria, NHI ≤1, and PHRI=0
- to assess the inter-rater agreement between the CAD system and human endoscopists
- the ability of the AI CAD system to stratify the risk of incurring prespecified clinical outcomes at 12 months.

**Statistical analysis**

The sample size was previously calculated for the PICaSSO multicentre study to observe a difference in correlation with histology between PICaSSO and MES.[9] Data was stored in REDCap and analysed with Matlab (R2021b, The Mathworks Inc, MA, 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 cut-off value to determine endoscopic remission/endoscopic 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 AUROC. Confidence intervals were calculated according to Clopper-Pearson [20] for sensitivity, specificity and accuracy, according to Mercaldo et al for PPV and NPV [reference 21] and by bootstrapping 1000 times the data and computing the 5 and 95 percentile of the bootstrapped sample for the AUROCs. The statistical differences in AUROC for different classifiers were computed using the non-parametric approach by DeLong et al.[reference 22] The agreement among human endoscopic assessments and AI-estimated outputs was measured with Cohen 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, hospitalisation, drug change or optimisation) within 12 months. Different survival curves and hazard ratios were computed for the groups obtained by endoscopist's PICaSSO and UCEIS scoring, and AI's VCE and WLE scoring.

The study was conducted and reported following the Checklist for artificial intelligence in Medical Imaging (CLAIM) criteria and the Checklist for Prediction Model Development and Validation (TRIPOD) (supplementary material)

**RESULTS**

Demographic characteristics of our study population are summarised in table 1.[9] Briefly, we included 283 patients, with an average age of 48.2 years (SD 14.8). Around 2/3 of patients were
in histologic remission depending on biopsy location and histologic score used. Supplementary Table 1.

Videos 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 with inadequate bowel preparation, the videos were divided into WLE-HD (539) and VCE (551) clips. In total 1090 clips comprised 638287 frames. 901 clips were rated as HQ, 189 LQ. 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 WLE-HD, 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 WLE-HD videos were available for the same patient and section, they were assigned to the same data set (training, validation, or testing) for better method comparison. The process is illustrated in Figure 1.

Primary Outcome

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

In the testing set, our system detected endoscopic remission/activity (PICaSSO ≤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 area under the ROC curve (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 UC activity in WLE-HD**

For the detection of endoscopic remission/activity in WLE-HD videos (UCEIS ≤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 HQ videos sub-analysis, 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 WLE-HD (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 endoscopy**
  Our CAD system, analysing the same VCE endoscopic videos, 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 restricting to HQ videos only. (Table 3)

- **Prediction of histological remission (RHI≤3; NHI≤1; PHRI=0) from WLE-HD endoscopy**
  AI prediction of HR with WLE-HD endoscopic videos 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 removing lower quality videos, the accuracy improved by 4-5%. (Supplementary Table 2)

- **Inter-rater agreement between AI and human endoscopists**
  The inter-rater agreement between the AI system and the human endoscopists in detecting endoscopic activity/remission, expressed as Cohen 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 WLE-HD. Given that the true value of the kappa coefficient lies within the confidence intervals with 95% probability, agreement in VCE videos is at least substantial, and at least fair in WLE-HD. (Table 2) In detecting histologic remission/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 and Supplementary Table 2)

- **AI assessment of risk of pre-specified 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 hospitalisation, colectomy, and UC treatment change due to relapse). Figure 3 presents the Kaplan-Meier curves for patients in remission or activity according to PICaSSO assessed by
human endoscopists (panel A) and the AI system (panel B). Panel A shows a strong association with risk of outcome for patient with activity (hazard ratio 4.59; 95% CI 1.88 - 11.2); panel B confirms how AI-assessed endoscopic activity was similarly associated with the same outcomes (hazard ratio 4.05; 95% CI 1.71 - 9.57). The same analysis obtained with WLE-HD classifying remission/activity according to 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 (Figure 3, panels D and E). 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) though the difference did not reach statistical significance. Supplementary figure 1.

DISCUSSION
The objective and reproducible evaluation of endoscopic activity is crucial to generalise assessment. VCE, through the PICaSSO score, has shown to bridge the discrepancy between traditional endoscopic and histologic evaluation, allowing the detection of subtle changes overlooked in conventional WL endoscopy,[23] regardless of the VCE platform.[24]

We developed the first CAD system to evaluate endoscopic and histologic activity and remission and predict clinical specified outcomes through VCE, in addition to conventional WLE-HD, thus 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%). Consistently with the hypothesis that VCE improves optical diagnosis, the same model had slightly worse diagnostic performance with WLE-HD (specificity 87% and sensitivity 72%). The statistical comparison of the two AUROCs supports this difference (p=0.02), although caution is necessary since the performance of the two models (VCE and WLE-HD) is assessed with different scores and cut-offs (PICaSSO≤3 for VCE and UCEIS≤1 for WLE-HD). We chose not to use the Mayo endoscopic score (MES) as it is not fully validated, its ER definition includes 0 or 1, and, as several studies show, its correlation with histology is lower than that of PICaSSO and UCEIS.[6,9]

In real-time, our CAD can provide an initial assessment of inflammation when using WLE-HD and 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. Trusting the AI-predicted endoscopic activity from VCE 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% considering only HQ videos. Of the 8 patients at risk of undertreatment 3 suffered a disease flare during follow up.

In the future, our system could successfully be implemented in both non-expert and expert clinical practice as well as clinical trials. When using the AI model to predict histology, the specificity remained strong (>80%), suggesting that the inflammatory activity seen in endoscopy corresponds to that found in histology. On the contrary, sensitivity ranged between 66 and 74% depending on the score, supporting the common notion that some features of histological inflammation are not visible with endoscopy. However, overall 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 WLE-HD has different possible explanations. First and foremost, VCE improves the detection of inflammation in human endoscopists, but there is no guarantee that an algorithm derives its predictions from the same mucosal features humans use. Secondly, even if it did, the system might still detect subtle changes also in WLE-HD 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 WLE-HD. Although different scores prevent a direct comparison, the results suggest that the assessment using VCE might be more reproducible.

Prediction of prognosis represents an exciting further step in the development of computer tools. The hazard ratios of suffering an adverse clinical outcome in the endoscopic remission and activity groups identified by human and VCE-AI point to an accurate stratification of risk of flare. The same classification using WLE-HD/UCEIS was slightly less robust though 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 data sets become available and the system is further refined.

Our work has several strengths. Firstly, to the best of our knowledge, it is the first AI model developed for the assessment of colonoscopy videos based on an optical enhancement system and also 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 video were recorded and endoscopic scores derived. This apparently simple shrewdness is seldom found in other works and reinforces our observation. 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 to others, notably the work of Takenaka et al.[14] but allows a practical use more similar to real-life clinical observation while avoiding the discontinuity and possible selection bias of assessing selected pictures. Moreover, the computerised analysis can take place in real time (see example videos) or later, on request, providing 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 homogenous 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 which and how data are presented (ie. if dye is only used in quiescent patients, the algorithm might predict remission from presence of dye rather than the mucosa). This applies also to aspects such as video capture, lighting and recording. The multicentre source of data (11 centres, 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. First, all procedures were carried out in tertiary centres by endoscopists experienced in IBD optical diagnosis, potentially less representative of ordinary care settings. Secondly, the dataset was limited to the rectum and the sigmoid. Nevertheless, given the distribution of UC, the absence of more proximal segments is unlikely to impact the model functioning. [26] Videos were of different quality and this may have affected the
diagnostic performance, unsurprisingly, in fact after removing lower quality clips, the model performance increased. In addition, the system has not yet been assessed on responsiveness to treatment. Finally, our model was developed and tested with videos recorded only with the iScan (Pentax, Tokyo, Japan) platform. We recently reported that PICaSSO is valid for other optical enhancement platforms.[24] Nevertheless, a prospective multicentre study to validate the system on other VCE platforms is planned.

In conclusion, we developed and tested an AI to distinguish endoscopic and histologic activity from remission of UC in colonoscopy videos of both WLE-HD and VCE. The CAD system developed on VCE videos showed a greater diagnostic performance for the assessment of endoscopic activity compared to the same system based on WLE-HD videos. This tool has multiple potential applications, such as standardising 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 most affected areas. Building on our previous work on computerised assessment of UC histopathology [19], we plan to integrate the two tools and further validate them in a large multicentre study.

References


References

Legends

**Figure 1. Study videos partition.** All endoscopy recordings were first edited to separate WLE and VCE parts, then videos were divided in three sets for training, validation, and testing of the AI models to detect ER/activity according to UCEIS and PICaSSO, predict HR as defined by NHI, RHI and PHRI, and predict future outcomes.

**Figure 2. Artificial Intelligence Architecture.** The classification stage of a pre-trained ResNet50 CNN classifier was trained to detect healing or active inflammation on video frames. Two separate networks were trained to detect ER/activity according UCEIS and PICaSSO from frames in WLE-HD and VCE videos, respectively. We present examples of endoscopic images showing features of ER and activity in WLE-HD and VCE that were used to train the model and examples of the AI outputs frames in WLE-HD and VCE.

**Figure 3 Clinical outcome prediction.** Kaplan-Meier survival functions for the two groups of patients (ER versus endoscopic activity) were estimated to evaluate the cumulative risk of incurring any of the specified adverse clinical outcomes (surgery, hospitalisation, drug change or optimisation) within 12 months as assessed by human endoscopists (panel A; panel C), by the AI predicted endoscopic score (panel B; D)

**Supplementary Figure 1** Distribution of AUROC values (obtained through bootstrapping) to predict any adverse outcome within 12 months for a patient, given their endoscopic assessment at baseline. The different assessment are: UCEIS (UCEIS score as assessed by the endoscopist on WLE-HD), AI UCEIS (probability of presence of active inflammation obtained by the AI trained on WLE-HD), PICaSSO (PICaSSO score as assessed by the endoscopist on VCE), AI PICaSSO (probability of presence of active inflammation obtained by the AI trained on VCE).

* means a statistical significant difference between two AUROCs with p<0.05

**Video:** Example of the AI system detection of endoscopy remission or activity in WLE-HD and VCE videos. All the AI outputs are provided in real-time.
### Supplementary Table 1  Distribution of histological activity

<table>
<thead>
<tr>
<th>RHI</th>
<th>Rectum</th>
<th>Sigmoid</th>
<th>NHI</th>
<th>Rectum</th>
<th>Sigmoid</th>
<th>PHRI</th>
<th>Rectum</th>
<th>Sigmoid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>N of Patients (%)</td>
<td>Score</td>
<td>N of patients (%)</td>
<td>Score</td>
<td>N of patients (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>98 (34.6)</td>
<td>100 (35.3)</td>
<td>0</td>
<td>167 (59.0)</td>
<td>170 (60.0)</td>
<td>0</td>
<td>197 (69.6)</td>
<td>206 (72.8)</td>
</tr>
<tr>
<td>1</td>
<td>68 (24.0)</td>
<td>67 (23.6)</td>
<td>1</td>
<td>31 (10.9)</td>
<td>36 (12.7)</td>
<td>1</td>
<td>80 (28.3)</td>
<td>72 (25.4)</td>
</tr>
<tr>
<td>2</td>
<td>30 (10.6)</td>
<td>36 (12.7)</td>
<td>2</td>
<td>27 (9.5)</td>
<td>28 (9.9)</td>
<td>Missing</td>
<td>6 (2.1)</td>
<td>5 (1.8)</td>
</tr>
<tr>
<td>3</td>
<td>10 (3.5)</td>
<td>15 (5.3)</td>
<td>3</td>
<td>31 (10.9)</td>
<td>30 (10.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7 (2.5)</td>
<td>5 (1.8)</td>
<td>4</td>
<td>21 (7.4)</td>
<td>15 (5.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10 (3.5)</td>
<td>8 (2.8)</td>
<td>Missing</td>
<td>6 (2.1)</td>
<td>4 (1.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5 (1.8)</td>
<td>6 (2.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8 (2.8)</td>
<td>5 (1.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8 (2.8)</td>
<td>8 (2.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>16 (5.6)</td>
<td>9 (3.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>8 (2.8)</td>
<td>11 (3.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5 (1.8)</td>
<td>6 (2.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3 (1.0)</td>
<td>3 (1.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>7 (2.5)</td>
<td>4 (1.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 2. Performance in prediction of histologic healing with WLE-HD

<table>
<thead>
<tr>
<th>Diagnostic performance</th>
<th>RHI ≤3 or &gt;3</th>
<th>NHI ≤1 or &gt;1</th>
<th>PHRI ≤1 or &gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Testing cohort</td>
<td>Testing cohort</td>
<td>Testing cohort</td>
</tr>
<tr>
<td></td>
<td>WLE-HD 242 videos</td>
<td>WLE-HD 183 HQ videos</td>
<td>WLE-HD 242 videos</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.66 (0.51-0.79)</td>
<td>0.72 (0.55-0.86)</td>
<td>0.67 (0.53-0.79)</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.84 (0.78-0.89)</td>
<td>0.86 (0.78-0.91)</td>
<td>0.86 (0.80-0.91)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.52 (0.42-0.61)</td>
<td>0.56 (0.45-0.67)</td>
<td>0.59 (0.49-0.68)</td>
</tr>
<tr>
<td>NPV</td>
<td>0.90 (0.86-0.93)</td>
<td>0.93 (0.88-0.96)</td>
<td>0.89 (0.85-0.92)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.80 (0.74-0.85)</td>
<td>0.84 (0.77-0.89)</td>
<td>0.81 (0.75-0.86)</td>
</tr>
<tr>
<td>Cohen Kappa</td>
<td>0.45 (0.31-0.69)</td>
<td>0.53 (0.38-0.68)</td>
<td>0.50 (0.37-0.63)</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.80 (0.72-0.88)</td>
<td>0.81 (0.73-0.88)</td>
<td>0.79 (0.72-0.87)</td>
</tr>
</tbody>
</table>

*and no neutrophils in the lamina propria or epithelium
### Table 1. Patients’ demographics and characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Patients (n=283)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y) mean ± sd</td>
<td>48.2 ± 14.8</td>
</tr>
<tr>
<td>Male sex n (%)</td>
<td>165 (58%)</td>
</tr>
<tr>
<td>Disease duration (y) mean ± sd</td>
<td>14.7 ± 10</td>
</tr>
<tr>
<td>Primary Sclerosing Cholangitis</td>
<td>37 (13%)</td>
</tr>
<tr>
<td><strong>Extension n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Left-sided colitis</td>
<td>122 (43.1%)</td>
</tr>
<tr>
<td>Sub-total or total colitis</td>
<td>159 (56.2%)</td>
</tr>
<tr>
<td>Missing data</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td><strong>Therapy in last 12 months n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>No treatment</td>
<td>15 (5.3%)</td>
</tr>
<tr>
<td>5-ASA</td>
<td>220 (77.7%)</td>
</tr>
<tr>
<td>Corticosteroids</td>
<td>71 (25.0%)</td>
</tr>
<tr>
<td>Immunomodulators</td>
<td>69 (24.4%)</td>
</tr>
<tr>
<td>Biologics</td>
<td>105 (37.1%)</td>
</tr>
<tr>
<td><strong>Mayo Endoscopic Score n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Mayo 0</td>
<td>156 (55.1%)</td>
</tr>
<tr>
<td>Mayo 1</td>
<td>46 (16.3%)</td>
</tr>
<tr>
<td>Mayo 2</td>
<td>52 (18.4%)</td>
</tr>
<tr>
<td>Mayo 3</td>
<td>27 (9.5%)</td>
</tr>
<tr>
<td>Missing data</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td><strong>UCEIS</strong></td>
<td></td>
</tr>
<tr>
<td>Remission (≤1)</td>
<td>200 (71%)</td>
</tr>
<tr>
<td>Active (&gt;1)</td>
<td>83 (29%)</td>
</tr>
<tr>
<td><strong>PICaSSO Score</strong></td>
<td></td>
</tr>
<tr>
<td>Remission (&lt;3)</td>
<td>191 (69%)</td>
</tr>
<tr>
<td>Active (≥3)</td>
<td>86 (31%)</td>
</tr>
<tr>
<td>Rectum</td>
<td>208 (73%)</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>75 (27%)</td>
</tr>
</tbody>
</table>

*missing data due to inadequate bowel preparation precluding the endoscopic scoring. These patients were not included in the overall analysis.*
Table 2. Diagnostic performance in prediction of endoscopic healing according to PICaSSO score with VCE, and to UCEIS with WLE-HD.

<table>
<thead>
<tr>
<th>Diagnostic performance</th>
<th>PICaSSO ≤3 or &gt;3</th>
<th>UCEIS ≤1 or &gt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>Testing</td>
<td>Validation</td>
</tr>
<tr>
<td>VCE 62 videos</td>
<td>VCE 244 videos</td>
<td>VCE 196 HQ videos</td>
</tr>
<tr>
<td>Sensitivity (95% CI)</td>
<td>0.89 (0.66-0.98)</td>
<td>0.79 (0.63-0.90)</td>
</tr>
<tr>
<td>Specificity (95% CI)</td>
<td>0.93 (0.81-0.99)</td>
<td>0.95 (0.91-0.98)</td>
</tr>
<tr>
<td>PPV (95% CI)</td>
<td>0.85 (0.65-0.94)</td>
<td>0.77 (0.64-0.86)</td>
</tr>
<tr>
<td>NPV (95% CI)</td>
<td>0.95 (0.84-0.99)</td>
<td>0.96 (0.92-0.97)</td>
</tr>
<tr>
<td>Accuracy (95% CI)</td>
<td>0.92 (0.82-0.97)</td>
<td>0.92 (0.88-0.95)</td>
</tr>
<tr>
<td>Cohen Kappa (95% CI)</td>
<td>0.81 (0.66-0.97)</td>
<td>0.73 (0.61-0.85)</td>
</tr>
<tr>
<td>AUROC (95% CI)</td>
<td>0.94 (0.91-0.97)</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Diagnostic performance in prediction of histologic healing with VCE

<table>
<thead>
<tr>
<th>Diagnostic performance</th>
<th>RHI ≤3* or &gt;3</th>
<th>NHI ≤1 or &gt;1</th>
<th>PHRI ≤1 or &gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Testing cohort</td>
<td>Testing cohort</td>
<td>Testing cohort</td>
</tr>
<tr>
<td></td>
<td>VCE 242 videos</td>
<td>VCE 193 HQ videos</td>
<td>VCE 242 videos</td>
</tr>
<tr>
<td>Sensitivity (95% CI)</td>
<td>0.73 (0.59-0.85)</td>
<td>0.74 (0.56-0.87)</td>
<td>0.65 (0.51-0.77)</td>
</tr>
<tr>
<td>Specificity (95% CI)</td>
<td>0.86 (0.80-0.91)</td>
<td>0.87 (0.81-0.92)</td>
<td>0.86 (0.80-0.91)</td>
</tr>
<tr>
<td>PPV (95% CI)</td>
<td>0.57 (0.47-0.66)</td>
<td>0.57 (0.44-0.66)</td>
<td>0.59 (0.49-0.68)</td>
</tr>
<tr>
<td>NPV (95% CI)</td>
<td>0.93 (0.89-0.95)</td>
<td>0.94 (0.90-0.96)</td>
<td>0.89 (0.85-0.92)</td>
</tr>
<tr>
<td>Accuracy (95% CI)</td>
<td>0.83 (0.78-0.88)</td>
<td>0.85 (0.79-0.90)</td>
<td>0.81 (0.75-0.86)</td>
</tr>
<tr>
<td>Cohen Kappa (95% CI)</td>
<td>0.54 (0.41-0.67)</td>
<td>0.54 (0.39-0.69)</td>
<td>0.49 (0.36-0.62)</td>
</tr>
<tr>
<td>AUROC (95% CI)</td>
<td>0.83 (0.75-0.90)</td>
<td>0.81 (0.74-0.88)</td>
<td>0.81 (0.74-0.88)</td>
</tr>
</tbody>
</table>

*and no neutrophils in the lamina propria or epithelium
AI Development

WLE Endo
Prediction of ER/activity (UCEIS)

Training 239 videos
Validation 58 videos
Test 242 videos

UCEIS-AI
Operating point
Predicting ER/activity

Predicting UCEIS

VCE Endo
Prediction of ER/activity (PICaSSO)

Training 245 videos
Validation 62 videos
Test 244 videos

PICaSSO-AI
Operating point
Predicting ER/activity

Predicting PICaSSO

Total patients 283
Total videos 1090
Total frames 67280