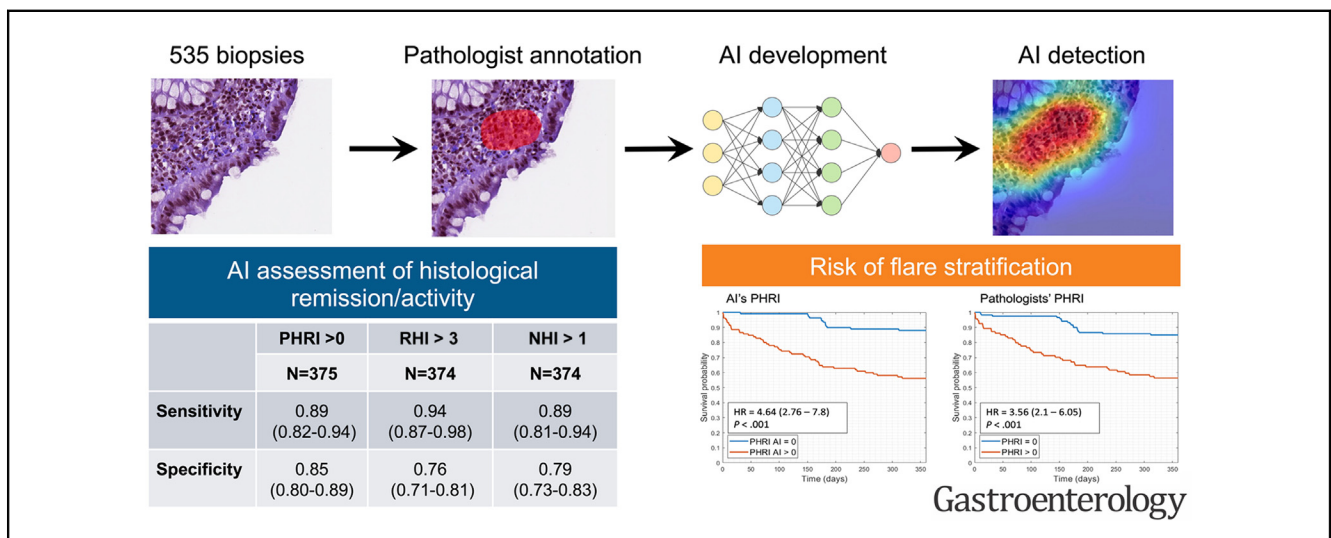# INFLAMMATORY BOWEL DISEASE

Artificial Intelligence Enabled Histological Prediction of Remission or Activity and Clinical Outcomes in Ulcerative Colitis

Marietta Iacucci,[1,2,3,4,*] **Tommaso Lorenzo Parigi,**[1,*] **Rocio Del Amor,**[5,*]
Pablo Meseguer,[5,*] Giulio Mandelli,[6] Anna Bozzola,[6] Alina Bazarova,[7] Pradeep Bhandari,[8]
Raf Bisschops,[9] Silvio Danese,[10,11] Gert De Hertogh,[12] Jose G. Ferraz,[13] Martin Goetz,[14]
Enrico Grisan,[15,16] Xianyong Gui,[17] Bu Hayee,[18] Ralf Kiesslich,[19] Mark Lazarev,[20]
Remo Panaccione,[13] Adolfo Parra-Blanco,[21,22] Luca Pastorelli,[23] Timo Rath,[24]
Elin S. Røyset,[25] Gian Eugenio Tontini,[26,27] Michael Vieth,[28] Davide Zardo,[29] Subrata Ghosh,[4]
Valery Naranjo,[5] and Vincenzo Villanacci[6]

[1]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, United Kingdom; [2]NIHR Wellcome Trust Clinical Research Facility, University Hospital Birmingham, Birmingham, United Kingdom; [3]Department of Gastroenterology, University Hospitals Birmingham NHS Trust, Birmingham, United Kingdom; [4]APC Microbiome Ireland, College of Medicine and Health, University College Cork, Cork, Ireland; [5]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Valencia, Spain; [6]Institute of Pathology, ASST Spedali Civili, University of Brescia, Brescia, Italy; [7]Institute for Biological Physics, University of Cologne, Cologne, Germany; [8]Department of Gastroenterology, Queen Alexandra Hospital, Portsmouth, United Kingdom; [9]Department of Gastroenterology, University Hospitals Leuven, Leuven, Belgium; [10]Gastroenterology and Endoscopy, IRCCS Ospedale San Raffaele, Milan, Italy; [11]University Vita-Salute San Raffaele, Milan, Italy; [12]Laboratory of Translational Cell and Tissue Research, Department of Imaging and Pathology, Faculty of Medicine, KU, Leuven, Belgium; [13]Division of Gastroenterology and Hepatology, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada; [14]Division of Gastroenterology, Klinikum Böblingen, Böblingen, Germany; [15]Department of Information Engineering, University of Padova, Padova, Italy; [16]School of Engineering, London South Bank University, London, United Kingdom; [17]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington; [18]King's Health Partners Institute of Therapeutic Endoscopy, King's College Hospital, London, United Kingdom; [19]Division of Gastroenterology, Helios HSK Wiesbaden, Wiesbaden, Germany; [20]Department of Gastroenterology, Johns Hopkins Hospital, Baltimore, Maryland; [21]NIHR Nottingham Biomedical Research Centre, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom; [22]Department of Gastroenterology, University of Nottingham, Nottingham, United Kingdom; [23]Department of Health Sciences, School of Medicine Ospedale San Paolo, Università degli Studi di Milano, Milan, Italy; [24]Department of Gastroenterology, Friedrich Alexander University of Erlangen, Nuremberg, Germany; [25]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway; [26]Fondazione IRCCS Ca'Granda Ospedale Maggiore Policlinico, Milan, Italy; [27]Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy; [28]Institute of Pathology, Friedrich Alexander University Erlangen-Nuremberg, Klinikum Bayreuth, Bayreuth, Germany; and [29]Department of Pathology, San Bortolo Hospital, Vicenza, Italy

| 535 biopsies | Pathologist annotation | AI development | AI detection |

| AI assessment of histological remission/activity | | | | Risk of flare stratification |

	PHRI >0	RHI > 3	NHI > 1
	N=375	N=374	N=374
Sensitivity	0.89 (0.82-0.94)	0.94 (0.87-0.98)	0.89 (0.81-0.94)
Specificity	0.85 (0.80-0.89)	0.76 (0.71-0.81)	0.79 (0.73-0.83)

AI's PHRI — HR = 4.64 (2.76 – 7.8), $P < .001$; PHRI AI = 0, PHRI AI > 0

Pathologists' PHRI — HR = 3.56 (2.1 – 6.05), $P < .001$; PHRI = 0, PHRI > 0

Gastroenterology

BACKGROUND & AIMS: Microscopic inflammation has significant prognostic value in ulcerative colitis (UC); however, its assessment is complex with high interobserver variability. We aimed to develop and validate an artificial intelligence (AI) computer-aided diagnosis system to evaluate UC biopsies and predict prognosis. **METHODS:** A total of 535 digitalized

biopsies (273 patients) were graded according to the PICaSSO Histologic Remission Index (PHRI), Robarts, and Nancy Histological Index. A convolutional neural network classifier was trained to distinguish remission from activity on a subset of 118 biopsies, calibrated on 42 and tested on 375. The model was additionally tested to predict the corresponding endoscopic assessment and occurrence of flares at 12 months. The system output was compared with human assessment. Diagnostic performance was reported as sensitivity, specificity, prognostic prediction through Kaplan-Meier, and hazard ratios of flares between active and remission groups. We externally validated the model in 154 biopsies (58 patients) with similar characteristics but more histologically active patients. **RESULTS:** The system distinguished histological activity/remission with sensitivity and specificity of 89% and 85% (PHRI), 94% and 76% (Robarts Histological Index), and 89% and 79% (Nancy Histological Index). The model predicted the corresponding endoscopic remission/activity with 79% and 82% accuracy for UC endoscopic index of severity and Paddington International virtual ChromoendoScopy ScOre, respectively. The hazard ratio for disease flare-up between histological activity/remission groups according to pathologist-assessed PHRI was 3.56, and 4.64 for AI-assessed PHRI. Both histology and outcome prediction were confirmed in the external validation cohort. **CONCLUSION:** We developed and validated an AI model that distinguishes histologic remission/activity in biopsies of UC and predicts flare-ups. This can expedite, standardize, and enhance histologic assessment in practice and trials.

Keywords: Ulcerative Colitis; PICaSSO Histologic Remission Index; Computer-aided Diagnosis; Convolutional Neural Network; Robarts Histopathology Index.

WHAT YOU NEED TO KNOW
BACKGROUND AND CONTEXT
Histologic remission is the optimal goal of treatment in ulcerative colitis; however, histologic assessment is limited by low agreement between pathologists.
NEW FINDINGS
Our validated artificial intelligence model accurately distinguishes remission/inflammation according to 3 histological indices and stratifies risk of flare similarly to human physicians.
LIMITATIONS
The system cannot distinguish between different grades of disease severity.
CLINICAL RESEARCH RELEVANCE
Our computer tool can speed up, simplify, and standardize histologic assessment of ulcerative colitis in clinical practice and clinical trials, and provide accurate prognostic information to the clinician.
BASIC RESEARCH RELEVANCE
Automated detection of neutrophils can help shed light on their role in mucosal inflammation. Similar systems can be developed for other cell types or tissues, expanding the fields of application.

Ulcerative colitis (UC) is a chronic inflammatory bowel disease (IBD) characterized by a remitting-relapsing course.[1] Treatment of UC aims to extinguish inflammation to prevent complications, and histopathology is the most stringent method to detect the presence of inflammation and distinguish it from remission. Several studies have shown that patchy microscopic disease activity, even in absence of endoscopic features, is associated with an increased risk of flare and hospitalization.[2] Consistently, histologic remission (HR) correlates with improved clinical outcomes and has become a target of treatment.[3] To assess disease activity, biopsies are routinely taken in different segments of the colon; however, grading severity remains difficult. More than 30 histological indices have been proposed, suggesting none are ideal, and their adoption in clinical practice remains modest.[4] Scoring is time-consuming, requires dedicated training, and, more importantly, is limited by high interobserver variability.[5,6] To overcome this, clinical trials resort to expensive centralized readings to attempt reliable measurements.

Computer-aided diagnosis (CAD) systems based on artificial intelligence (AI) are increasingly used to simplify and standardize the evaluation of medical imaging. Successful applications in digital pathology include the quantification of the expression of molecular targets, such as hormone receptors and HER2 in breast cancer, or protein Ki67 in carcinoid tumors,[7] automated morphological analysis of nuclei,[8] and cellular features.[9] These technologies hold promise to enhance assessment, simplify interpretation, and resolve discrepancies between pathologists. To the best of our knowledge, in the field of UC pathology, only 2 AI models have been developed, the first, by Vande Casteele and colleagues,[10] focused on the detection of eosinophils and their correlation with disease activity; the second, by our group, concentrated on neutrophils as hallmarks of activity.[11]

We now present a comprehensive study of digital pathology computerized image analysis with a new and improved model to detect UC disease activity as defined by different histologic indices: Paddington International virtual ChromoendoScopy ScOre (PICaSSO) Histologic Remission Index (PHRI), Robarts Histologic Index (RHI), and Nancy Histologic Index (NHI). In addition, we used this AI-enabled

---

model to forecast the disease flare-up indicated by pre-specified clinical outcomes, and replicated our results in a separate external cohort.

Methods

Patients

For our main analysis, patients were recruited from 11 international centers between September 2016 and November 2019. Digitalized biopsies were not available for 1 center, so the study was carried out on data from 10 centers. The inclusion criteria were an established diagnosis of UC for more than 1 year regardless of disease activity and an indication to undergo colonoscopy. Exclusion criteria were contraindications to the procedure or biopsies, inability to provide consent, and inadequate bowel preparation. The study was approved by the central research ethics committee (17/WM/0223) for the United Kingdom centers and the local responsible ethics committees for each international center. To provide external validation, we collected a second cohort of patients from 2 centers, Birmingham in the United Kingdom and Brescia in Italy, with the same inclusion and exclusion criteria, who underwent colonoscopy with at least 1 biopsy in the rectum and in the sigmoid, and who were followed up for at least 12 months. Ethics approval for the external cohorts were Ref. 17/NI/0148 (Birmingham) and NP 5126 – STUDIO MICI-AI (Brescia). The study was approved by the West Midlands Research Ethics Committee (17/WM/0223). All patients gave informed consent to participate in the study.

Digital Pathology

At least 2 targeted tissue samples were taken in the sigmoid and rectum from the most representative areas of inflammation or healing, the same areas where the endoscopic assessment was recorded. Samples from the same segment were placed together in a single glass and processed as 1 biopsy. All samples were fixed in formalin, stained with hematoxylin and eosin, and digitally scanned at ×40 (0.25 μm per pixel) using Aperio Digital Pathology Scanning system (Leica Biosystem). The biopsies from the original cohort were assessed by 1 of the 6 expert IBD pathologists (Gert De Hertogh, Xianyong Gui, Elin S. Røyset, Michael Vieth, Vincenzo Villanacci, Davide Zardo) blinded to clinical and endoscopic information. In the external validation cohort, biopsies were assessed by either Vincenzo Villanacci or Davide Zardo. Histological activity was graded according to RHI, NHI, and the newly developed PHRI. The cutoffs for HR were the following: RHI ≤3 without neutrophils in the epithelium or lamina propria,[12] NHI ≤ 1,[13] and PHRI = 0.[11]

The subset of biopsies used for model training were digitally annotated by 3 expert gastrointestinal pathologists (Vincenzo Villanacci, Giulio Mandelli, Alina Bazarova). This was carried out by circling the areas with and without neutrophils using MicroDraw (NAAT).

Endoscopic Assessment

All procedures were performed with high-definition scopes (7010 processor, HiLine series colonoscopes; Pentax, Tokyo, Japan). The colonic mucosa was first assessed in white light high-definition and scored using Mayo endoscopic score[14] and UC endoscopic index of severity (UCEIS),[15] then in virtual chromoendoscopy (iSCAN1, iSCAN2, and iSCAN3) and scored by the PICaSSO score.[16,17] Endoscopic remission was defined as UCEIS ≤1 or PICaSSO ≤3.

Clinical Outcomes

For prognosis analysis, the clinical outcomes of UC-related hospitalization, UC-related surgery, and initiation increase or changes in UC therapy, including steroids, immunomodulators, and biological agents, driven by worsening symptoms, were chosen as proxies for disease flare and recorded at follow-up phone calls or visits 12 months after endoscopy in the initial cohort, or up to 33 months in the external validation cohort.[11,16]

CAD System

To implement PHRI in a CAD system, we designed a novel weakly supervised framework based on a multiple instance learning with constraints. Digitized biopsies (whole slide images [WSIs]) were down-sampled to ×20 resolution, divided into 512 × 512 × 3 areas (patches) with a 50% overlap, and patches with less than 20% of tissue were excluded. Then, a convolutional neural network based on a VGG16 architecture was applied to extract relevant features from each patch of the training biopsies. A separate module, based on a Squeeze and Excitation network, refined the low-dimensional features.[18] As a novelty, an attention-constrained module was added to focus the machine on the pixel-level annotations of the neutrophils' localization. The final whole-biopsy prediction was obtained through a multiple instance learning approach that weighted the assessment of each patch of the biopsy and aggregated them into a final binary result, remission or activity according to PHRI.[19] We hypothesized that neutrophils assessment, hence PHRI predictions, would align with other common indices, NHI and RHI, because both these scores heavily rely on the presence of neutrophils to define activity/remission. Therefore, without retraining the system, we compared the human scoring of NHI and RHI with the remission/activity classification of the CAD system to test our hypothesis (Figures 1 and 2). Details of model development are available in the Supplementary Material.

Outcome Measures and Statistical Analysis

The diagnostic performance of the CAD for the classification of UC remission/activity according to the respective histologic and endoscopic cutoffs was reported as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-Score, accuracy, and area under the receiver operating characteristic curve (AUROC). AUROCs were compared with the DeLong test. The 95% confidence intervals (CIs) for each metric were calculated. Of note, the presence of UC activity was considered the positive class in the binary classifications.

For the prognosis analysis, we investigated the cumulative risk of incurring any prespecified adverse clinical outcome (UC-related hospitalization, UC-related surgery, and initiation increase or changes in UC therapy due to inflammatory activity) during follow-up. The Cox proportional hazard model was used to compute Kaplan-Meier survival curves and calculate hazard ratios for the remission and active groups obtained by human pathologists scoring (PHRI, RHI, and NHI), AI-predicted PHRI scoring, and human endoscopists scoring (UCEIS and PICaSSO).
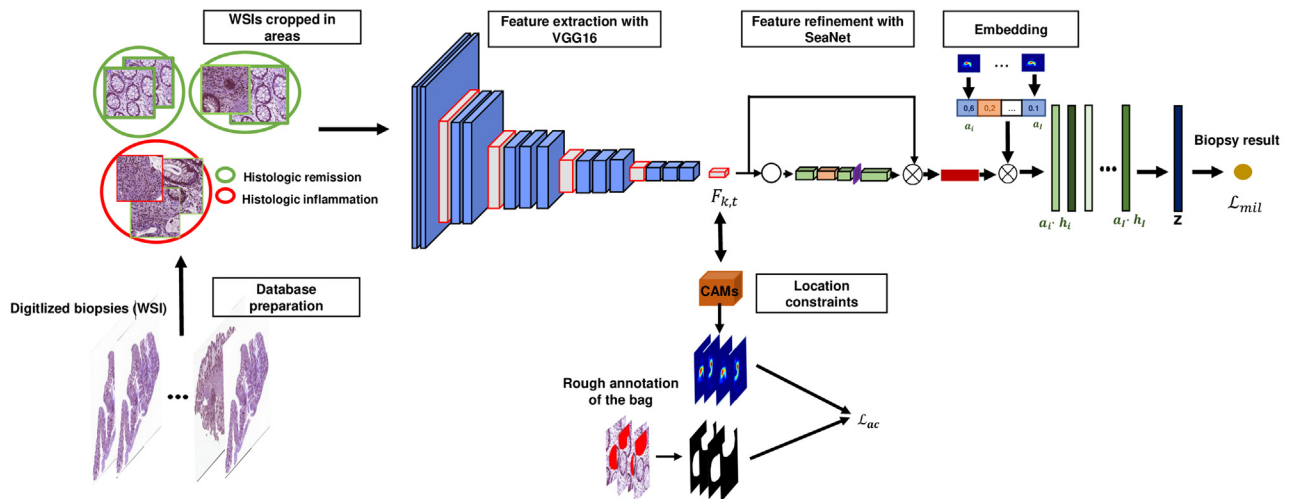
Figure 1. Artificial intelligence model. Framework of the deep learning approach to detect UC activity. (1) Biopsies are digitalized into WSIs. (2) A training set of WSIs is labeled by the pathologist as active or in remission. The WSIs are then divided in areas (patches) for computational reasons. (3) A feature extraction model (VGG16) is trained to recognize the feature (neutrophils) associated with activity. (4) An additional attention-constraint module is implemented to focus the machine on neutrophils thanks to the detailed pixel-level annotation of the WSI. (5) The assessment (vector) of each patch is weighted and combined with the other patches from the same WSI in a final aggregated result.

Data analysis was performed with Python (3.8) and MATLAB (Mathworks Inc).

Results

Initially, 535 biopsies were used to develop and test the model. These were collected in 273 patients, 40.7% were female, with an average age of 48.1 years (SD 14.8). Between two-thirds and three-fourths of biopsies were in histologic remission, depending on the score used to assess them (62% with PHRI, 76% with RHI, and 71% with NHI). Variability between pathologists was assessed on the same set of slides in our previous study,[11] and intraclass correlation coefficients were found to be
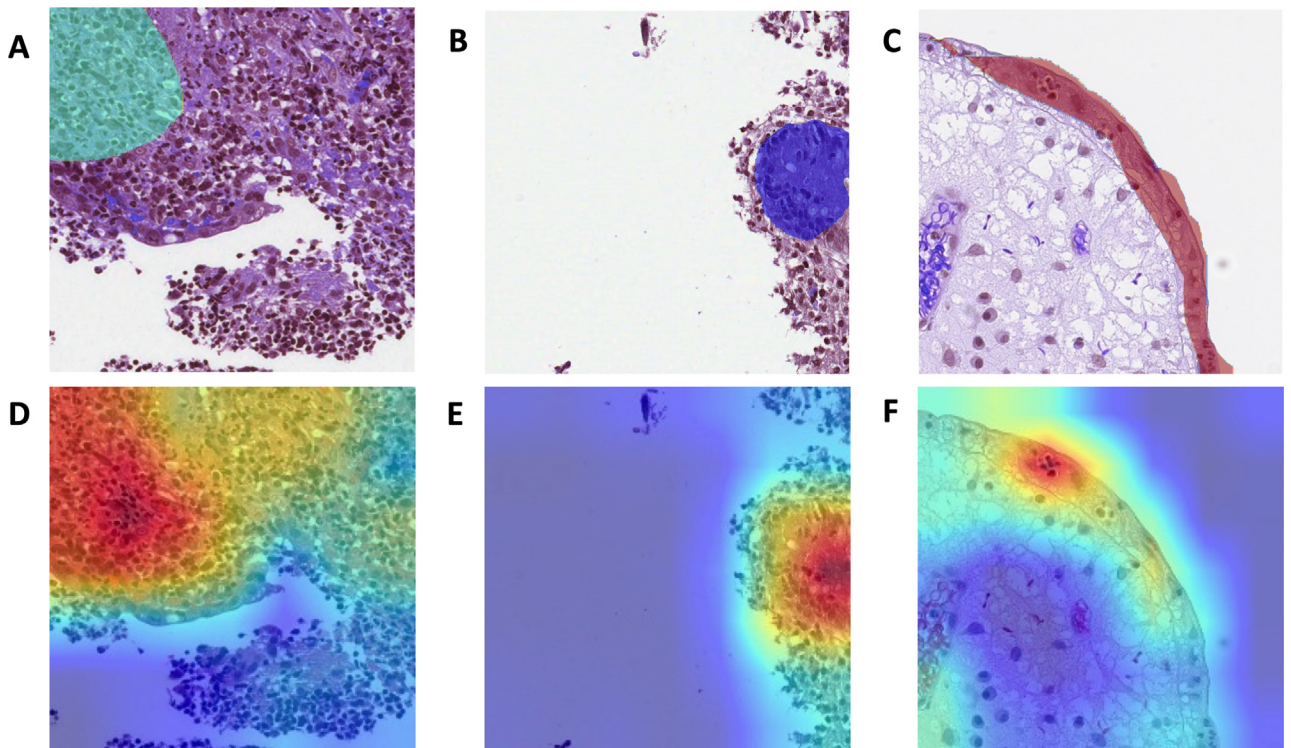
Figure 2. Examples of the CAD system's output. (A), (B), and (C): examples of biopsy annotations used to train the CAD system. (D), (E), and (F): examples of CAD system output. In the areas in *yellow* and *red* the system detects neutrophils. Importantly, the system recognizes neutrophils also where they were not previously annotated by the pathologist.

statistically similar for the 3 scores in a range between 0.77 and 0.85.

Of the initial 535 biopsies, 118 were used to train the model, 42 to calibrate it and 375 to test it. Subsequently, for the external validation, 154 additional biopsies from 58 patients with UC were used. The external cohort had roughly similar demographic characteristics, but twice the percentage of histologically active patients. Detailed characteristics are presented in Table 1.

CAD System Detection of Histological Activity/ Remission According to PHRI, RHI, and NHI

In the testing set of 375 biopsies, the CAD system distinguished histologic remission from disease activity defined according to PHRI ($=0$ vs ≥ 1) with a sensitivity of 89% (95% CI, 0.84–0.94), a specificity of 85% (95% CI, 0.80–0.89), a PPV of 75% (95% CI, 0.69–0.80), an NPV of 94% (95% CI, 0.90–0.96), an accuracy of 87% (95% CI, 0.83–0.90), and an AUROC of 87% (95% CI, 0.83–0.90).

We then tested the same system trained to detect neutrophils and predict PHRI against human assessment of remission/activity according to RHI (>3 or neutrophils in the epithelium or lamina propria) and NHI (>1). The system differentiated histologic remission/activity according to RHI with a sensitivity of 94% (95% CI, 0.87–0.98), a specificity of 76% (95% CI, 0.71–0.81), a PPV of 53% (95% CI, 0.48–0.58), an NPV of 98% (95% CI, 0.95–0.99), an accuracy of 80% (95% CI, 0.76–0.84), and an AUROC of 85% (95% CI, 0.82–0.89). When tested on NHI's cutoff of activity/remission, the CAD system's sensitivity was 89% (95% CI, 0.81–0.94), specificity 79% (95% CI, 0.73–0.83), PPV 60% (95% CI, 0.54–0.65), NPV 95% (95% CI, 0.92–0.97), accuracy 81% (95% CI, 0.77–0.85) and the AUROC 86% (95% CI, 0.83–0.90) (Table 2). Differences between the AUROCs were not statistically significant (PHRI vs RHI $P = .15$; PHRI vs NHI $P = .76$).

In the external validation cohort, the AUROC was 90% (95% CI, 0.86–0.95), with a sensitivity of 92% (95% CI, 0.86–0.96) and a specificity of 81% (95% CI, 0.63–0.93) (Table 2).

The CAD system delivered the results in an average of 9.8 seconds per slide.

CAD System Prediction of Endoscopic Activity/ Remission According to PICaSSO and UCEIS

The AI model predicted the corresponding endoscopic activity (UCEIS >1 and PICaSSO >3) with 78% (95% CI, 0.70–0.85) and 86% (95% CI, 0.77–0.92) sensitivity, 80% (95% CI, 0.74–0.84) and 78% (95% CI, 0.73–0.83) specificity, and an AUROC of 79% (95% CI, 0.75–0.83) and 82% (95% CI, 0.78–0.86), respectively, for the 2 scores (Supplementary Table 1).

CAD System Prognosis Prediction

Grouping patients in histologic remission or activity based on pathologists' assessment, the hazard ratio between the 2 groups for suffering any prespecified adverse clinical event, a proxy for flare-up, was 3.56 (95% CI, 2.10–6.05)

Table 1. Demographics Characteristics

Characteristics		
	Original cohort	Validation cohort
Number of biopsies	535	154
Number of patients	273	58
Age, y, mean ± SD	48.1 ± 14.8	44 ± 16
Female, n (%)	111 (40.7)	25 (43)
Disease duration, mean ± SD	14.6 ± 12.2	11.2 ± 9.3
Extension, n (%)		
Left-sided colitis	116 (42.5)	23 (39.6)
Sub-total colitis or total colitis	153 (56.0)	35 (60.4)
Missing data	4 (1.5)	0 (0)
Therapy in previous 12 months, n (%)		
No treatment	14 (5.1)	1 (1.7)
5-ASA	205 (75.1)	47 (81.0)
Corticosteroids	64 (23.4)	25 (43.1)
Immunomodulators	64 (23.4)	14 (24.1)
Biologics	103 (37.7)	16 (27.6)
Endoscopic activity		
Mayo Endoscopic Score, n (%)		
Mayo 0	143 (52.4)	11 (19.0)
Mayo 1	45 (16.5)	11 (19.0)
Mayo 2	53 (19.4)	23 (39.6)
Mayo 3	29 (10.6)	13 (22.4)
Missing data	3 (1.1)	0 (0)
UCEIS[a]		
Remission (≤ 1)	371 (67.9)	48 (31.2)
Active (>1)	172 (31.5)	103 (66.9)
Missing data	3 (0.6)	3 (1.9)
PICaSSO[a]		
Remission (≤ 3)	418 (76.6)	Not available
Active (>3)	126 (23.1)	
Missing data	2 (0.3)	
Histology		
PHRI[a]		
Remission ($=0$)	342 (62.6)	33 (21.4)
Activity (≥ 1)	200 (36.6)	121 (78.6)
Missing Data	4 (0.7)	0 (0)
RHI[a]		
Remission (≤ 3 + no neutrophils)[b]	413 (75.6)	34 (22.1)
Activity (>3)	128 (23.4)	115 (74.7)
Missing data	5 (0.9)	5 (3.2)
NHI[a]		
Remission (≤ 1)	389 (71.2)	27 (17.5)
Activity (>1)	152 (27.8)	122 (79.2)
Missing data	5 (0.9)	5 (3.2)

[a]UCEIS, PICaSSO, PHRI, RHI, and NHI figures refer to biopsies (or site of biopsy), not patient numbers.
[b]No neutrophils in the lamina propria and epithelium.

when classified according to PHRI, 4.28 (95% CI, 2.33–7.84) according to RHI, and 3.55 (95% CI, 2.03–6.23) according to NHI. When the same analysis was performed by the CAD system trained to distinguish PHRI activity/remission, the hazard ratio was 4.64 (95% CI, 2.76–7.8), similar to,

and numerically higher than, the corresponding analysis by human experts with any of the scores considered (Figure 3).

These results were confirmed in the external validation cohort where the AI-predicted classification of activity/ remission resulted in an HR of 2.241 (95% CI, 1.08–4.67) and of 2.591 (95% CI, 1.20–5.29) for the classification according to PHRI by human pathologists (Supplementary Figure 1).

Discussion

We present the results of an advanced AI-based CAD system able to analyze digitized biopsies to detect UC disease activity, as defined by multiple histologic scores, estimate the corresponding endoscopic activity, and predict future clinical outcomes. Our model was developed on digitalized WSIs of UC and trained on a new scoring index, PHRI, designed ad hoc to be implementable into machine learning models. PHRI defines activity and remission of UC according to the presence or absence of neutrophils in areas of the biopsy: superficial epithelium, lamina propria, cryptal epithelium, and cryptal lumen; absence of neutrophils from all areas (PHRI = 0) is considered remission, whereas their presence defines disease activity. In this study, the CAD system trained and tested on a large set of digitalized biopsies had a strong diagnostic performance to detect disease activity (PHRI >0) with an overall AUROC of 0.87, a sensitivity of 89%, and a specificity of 84%. These results were externally confirmed in an independent validation cohort. Despite the different mix of severity grades, the model maintained a good diagnostic performance, proving its applicability outside the original development setting. Histologic indices, RHI and NHI, were different from the model it was trained for, PHRI. However, the sensitivity of the CAD system for both RHI and NHI histologic remission/ activity was high (94% and 89%, respectively), although, admittedly, the PPVs were more modest. Of course, the prime role of neutrophils in defining histologic activity is also true in RHI and NHI,[12,13] which are endorsed for use in clinical trials by scientific societies, such as European Crohn's and Colitis Organisation.[20,21] Our tool represents an important first step toward a greater expedition and standardization of histologic reading in clinical trials.

The secondary analysis on prediction of endoscopic assessment demonstrated that the CAD system could predict the presence of endoscopic inflammation in the same area where the biopsies were taken with approximately 80% accuracy. Although imperfect, this result is consistent with human-assessed correlation between endoscopy and histology,[2] and therefore acceptable for a computer.

The most innovative application of our system is outcome stratification. The problem of uncertain prognosis is central to UC given its relapsing-remitting course and the variable response to treatment. We developed the first AI tool to stratify risk of flare based on histologic data, considered the most stringent assessment of UC activity, and observed a strong association between histologic activity and disease flare regardless of how the classification was made, by pathologists or by the CAD system. The hazard ratios, which express the strength of the association, were similar between the AI system and humans, demonstrating the ability of the computer to stratify the risk of flare comparably well to pathologists. In the external validation cohort, we confirmed the prognostic value of the model, which again identified patients at risk of flare similarly to humans.

In other words, we externally validated both the cross-sectional histologic assessment as well as the longitudinal prognostic stratification of the AI. The external cohort included a higher proportion of patients with active disease, but this did not influence the system prediction and hence further supports the use of our tool.

Our work has several strengths. First, the robustness of the data gathered as part of a prospective study, designed to match biopsies and endoscopic reading from the same colonic areas, is a prerequisite to investigating their

Table 2. CAD System Diagnostic Performance for Detecting Histological Activity/ Remission According to PHRI, RHI, and NHI

	CAD system diagnostic performance for histologic remission/activity				
	PHRI (PHRI >0)		RHI (RHI >3)[a]	NHI (NHI >1)	PHRI (PHRI >0)
	Calibration (n = 42)	Test (n = 375)	Test (n = 374)	Test (n = 374)	External validation (n = 154)
Sensitivity	0.76 (0.50–0.93)	0.89 (0.82–0.94)	0.94 (0.87–0.98)	0.89 (0.81–0.94)	0.92 (0.86–0.96)
Specificity	0.96 (0.80–0.99)	0.85 (0.80–0.89)	0.76 (0.71–0.81)	0.79 (0.73–0.83)	0.81 (0.63–0.93)
PPV	0.93 (0.65–0.99)	0.75 (0.69–0.80)	0.53 (0.48–0.58)	0.60 (0.54–0.65)	0.95 (0.90–0.98)
NPV	0.86 (0.72–0.93)	0.94 (0.90–0.96)	0.98 (0.95–0.99)	0.95 (0.92–0.97)	0.71 (0.57–0.82)
F1 Score	0.84 (0.73–0.94)	0.84 (0.80–0.88)	0.68 (0.63–0.73)	0.72 (0.67–0.76)	0.93 (0.90–0.97)
Accuracy	0.88 (0.74–0.96)	0.87 (0.83–0.90)	0.80 (0.76–0.84)	0.81 (0.77–0.85)	0.90 (0.84–0.94)
AUROC	0.86 (0.76–0.97)	0.87 (0.83–0.90)	0.85 (0.82–0.89)	0.86 (0.83–0.90)	0.90 (0.86–0.95)

AUROC, area under the receiver operator characteristic curve; CAD, computer aided diagnosis; NHI, Nancy Histological Index; NPV, negative predictive value; PHRI, PICaSSO Histological Remission Index; PPV, positive predictive value; RHI, Robarts Histological Index.
[a]No neutrophils in the lamina propria and epithelium.

INFLAMMATORY BOWEL DISEASE

INFLAMMATORY
BOWEL DISEASE
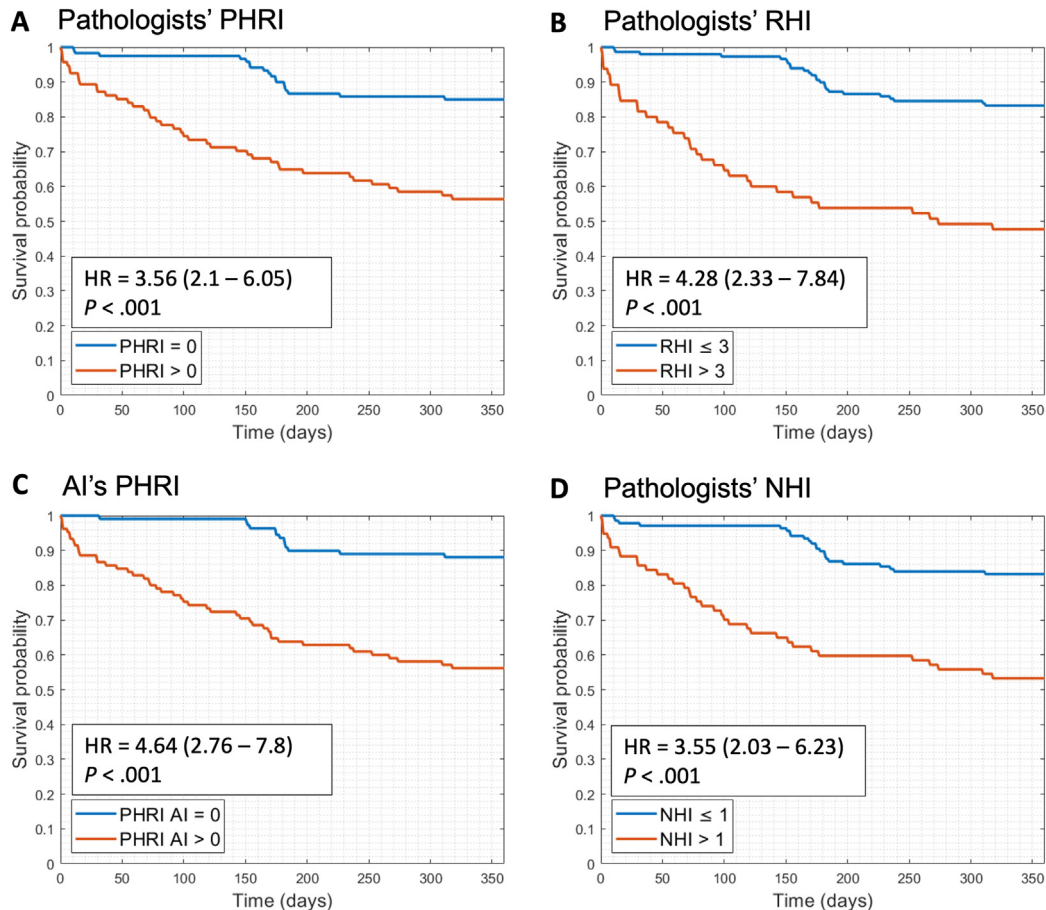
Figure 3. Kaplan-Meier curves of clinical events in histologic remission/activity groups. The Kaplan-Meier curves show the cumulative risk of incurring any of the specified adverse clinical outcomes (UC-related surgery, UC-related hospitalization, UC-treatment dose optimization or medication initiation due to inflammation) within 12 months after biopsy. The classification in histologic remission or activity was made by a pathologist grading inflammation according to PHRI (*A*), RHI (*B*), and NHI (*D*). (*C*) The same classification in PHRI activity/remission by the CAD system is shown. The hazard ratios express the increased risk of adverse events in the histologically active groups compared with the remission groups. Higher hazard ratios correspond to higher risk of event and better outcome stratification.

correlation. Moreover, prospective detailed follow-up adds to the wealth of data. This allows for assessing the immediate clinical implication of the model, providing practical valuable information to the clinician. The multicenter international collaboration, involving centers in different countries, allowed not only to access a large sample size but, equally importantly, to mitigate the risk of algorithm overfitting. Overfitting is the underperformance of an algorithm when applied to a set of unseen data different from those it was trained with, thus defeating its purpose. The main cause of overfitting is data homogeneity. In the field of histopathology, biopsy taking and processing involve several technical steps that can influence homogeneity, such as using the same type of biopsy forceps, tissue orientation, biopsy stains, glasses, and so forth. In our study, biopsies were collected and processed in the respective hospitals, adding heterogeneity, reducing the risk of overfitting, and supporting the generalizability of the results. To provide equal conditions between human and the AI and not introduce selection bias, we did not exclude biopsies with artifacts or lower quality as long as they had been considered sufficient

by the pathologist. We were aware of the generalizability problem of AI systems and therefore intended to address it prioritizing the heterogeneity of the dataset. Consistently with this, performance in the external validation cohort remained roughly similar supporting overall generalizability. In future prospective studies we will consider a subanalysis based on biopsy quality.

As pointed out before, the variability among scores is a major obstacle to interpreting and comparing histologic data. For this reason, we tested our system also on the 2 indices, RHI and NHI, recommended by scientific societies,[21] in an effort to independently compare the system and demonstrate its validity regardless of the score.

The main limitation of our work is that, at the current stage, the system cannot grade inflammatory activity. However, arguably, histologic disease assessment has its main role in detecting the persistence of inflammation when endoscopy is negative or mildly active. In the opposite case, when endoscopy already shows activity, histologic confirmation adds less prognostic information. Second, our system does not address dysplasia detection, the other main indication for

biopsies in UC. However, in the future we aim to expand our model to provide also dysplasia assessment alongside inflammation. Finally, our computer tool can be used only with digitalized biopsies, and although digitalized pathology is increasingly adopted, it is not widely available yet.

This study presents a CAD model for assisted diagnostic scoring of UC according to 3 scoring systems, a task until now prerogative of expert pathologists. We believe this tool will have an impact on both clinical trials and daily practice. In the latter, histologic reporting is still largely descriptive and nonstandard, thus would greatly benefit from a quick and objective assessment. Similarly, clinical trials in UC could efficiently overcome costly central readings. We are planning to conduct a prospective study in the context of therapeutic intervention using a targeted biopsy protocol, as current trials do not necessarily match biopsies from endoscopic evaluated areas.

In conclusion, our CAD system in real time accurately distinguished disease remission from activity as defined by the PHRI, RHI, and NHI and provided a good prediction of the corresponding endoscopic activity and the risk of flare. These results were confirmed in an external validation cohort. Future directions are to include dysplasia detection and to combine histologic and endoscopic AI models into an integrated tool to further improve disease monitoring and prediction.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2023.02.031.

References

1. Ungaro R, Mehandru S, Allen PB, et al. Ulcerative colitis. Lancet 2017;389:1756–1770.
2. 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.
3. Turner D, Ricciuto A, Lewis A, D'Amico, 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.
4. Mosli MH, Feagan BG, Sandborn WJ, et al. Histologic evaluation of ulcerative colitis: a systematic review of disease activity indices. Inflamm Bowel Dis 2014;20:564–575.
5. Mosli MH, Feagan BG, Zou G, et al. Reproducibility of histological assessments of disease activity in UC. Gut 2015;64:1765–1773.
6. Römkens TEH, Kranenburg P, Tilburg A van, et al. Assessment of histological remission in ulcerative colitis: discrepancies between daily practice and expert opinion. J Crohns Colitis 2018;12:425–431.
7. Niazi MKK, Parwani AV, Gurcan M. Digital pathology and artificial intelligence. Lancet Oncol 2019;20:e253–e261.
8. Durkee MS, Abraham R, Clark MR, et al. Artificial intelligence and cellular segmentation in tissue microscopy images. Am J Pathol 2021;191:1693–1701.
9. Brunt EM, Clouston AD, Goodman Z, et al. Complexity of ballooned hepatocyte feature recognition: defining a training atlas for artificial intelligence-based imaging in NAFLD. J Hepatol 2022;76:1030–1041.
10. Vande Casteele N, Leighton JA, Pasha SF, et al. Utilizing deep learning to analyze whole slide images of colonic biopsies for associations between eosinophil density and clinicopathologic features in active ulcerative colitis. Inflamm Bowel Dis 2022;28:539–546.
11. 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.
12. Pai RK, Khanna R, D'Haens GR, et al. Definitions of response and remission for the Robarts Histopathology Index. Gut 2019;68:2101–2102.
13. 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.
14. Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. N Engl J Med 1987;317:1625–1629.
15. 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.
16. 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.
17. 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.
18. Del Amor R, Launet L, Colomer A, et al. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. Artif Intell Med 2021;121:102197.
19. Del Amor R, Meseguer P, Parigi TL, et al. Constrained multiple instance learning for ulcerative colitis prediction using histological images. Comput Methods Programs Biomed 2022;224:107012.
20. Adamina M, Feakins R, Iacucci M, et al. ECCO topical review optimising reporting in surgery, endoscopy, and histopathology. J Crohns Colitis 2021;15:1089–1105.
21. Magro F, Doherty G, Peyrin-Biroulet L, et al. ECCO position paper: harmonization of the approach to ulcerative colitis histopathology. J Crohns Colitis 2020;14:1503–1511.

INFLAMMATORY BOWEL DISEASE

Correspondence
Address correspondence to Marietta Iacucci, MD, PhD, FASGE, AGAF, Department of Medicine, University College Cork, Clinical Sciences Building, Cork University Hospital, Cork T12EC8P, Ireland. e-mail: MIacucci@ucc.ie; iacuccim@yahoo.it.

CRediT Authorship Contributions
Marietta Iacucci, MD, PhD (Conceptualization: Lead; Data curation: Lead; Supervision: Lead; Writing – review & editing: Lead).

Tommaso Lorenzo Parigi, MD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Lead; Investigation: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Rocio Del Amor, MS (Formal analysis: Equal; Investigation: Equal; Software: Lead; Writing – review & editing: Equal).

Pablo Meseguer, MS (Formal analysis: Equal; Software: Equal; Writing – original draft: Equal; Writing – review & editing: Equal).

Giulio Mandelli, MD (Data curation: Equal).

Anna Bozzola, MD (Data curation: Equal).

Alina Bazarova, PhD (Methodology: Supporting; Writing – review & editing: Supporting).

Pradeep Bhandari, MD, PhD (Data curation: Equal).

Raf Bisschops, MD, PhD (Data curation: Equal).

Silvio Danese, MD, PhD (Writing – review & editing: Equal).

Gert De Hertogh, MD, PhD (Data curation: Equal).

Jose G Ferraz, MD, PhD (Data curation: Equal).

Martin Goetz, MD, PhD (Data curation: Equal).

Enrico Grisan, PhD (Methodology: Supporting; Software: Supporting).

Xianyong Gui, MD (Data curation: Equal).

Bu Hayee, MD, PhD (Data curation: Equal).

Ralf Kiesslich, MD, PhD (Data curation: Equal).

Mark Lazarev, MD, PhD (Data curation: Equal).

Remo Panaccione, MD, PhD (Data curation: Equal).

Adolfo Parra-Blanco, MD, PhD (Data curation: Equal).

Luca Pastorelli, MD, PhD (Data curation: Equal).

Timo Rath, MD, PhD (Data curation: Equal).

Elin Synnøve Røyset, MD, PhD (Data curation: Equal).

Gian Eugenio Tontini, MD, PhD (Data curation: Equal).

Michael Vieth, MD, PhD (Data curation: Equal).

Davide Zardo, MD, PhD (Data curation: Equal).

Subrata Ghosh, MD, PhD (Data curation: Equal; Supervision: Equal; Writing – review & editing: Equal).

Valery Naranjo, PhD (Formal analysis: Equal; Methodology: Equal; Software: Lead; Supervision: Equal).

Vincenzo Villanacci, MD, PhD (Data curation: Lead; Formal analysis: Equal; Supervision: Equal; Writing – review & editing: Equal).

INFLAMMATORY BOWEL DISEASE

Supplementary Material

Artificial Intelligence Appendix

A total of 535 biopsies of UC were analyzed and graded according to PHRI as in remission or activity by the expert pathologist assigned to each study center (6 pathologists split the biopsies collected in 11 centers). In a training subset of biopsies, single neutrophils were annotated in the 4 regions of the biopsy (lamina propria, surface epithelium, cryptal epithelium, and cryptal lumen) using an in-house software (MicroDraw). WSIs were then down-sampled to ×20 resolution, divided into 512 × 512 × 3 patches with a 50% overlap, and patches with less than 20% of tissue were excluded.

We then designed a novel convolutional neural network (CNN) that incorporates a backbone with attention constraints and a multiple instance learning (MIL) attention embedding. The model extracts a refined low-dimensionality representation using a feature extractor backbone (based on VGG16 architecture) and a feature-refinement module (based on a Squeeze and Excitation network).[1] This model also focuses on neutrophils at patch level due to the attention constraints that obtain a high-dimensionality activation map from the patch level feature. Finally, it is compared with the pixel-level annotations of the neutrophils to force the CNN to learn specific features for the single cells. After that, the extracted patch-level information is weighted with an index extracted from the class activation map and the results are combined to establish the presence 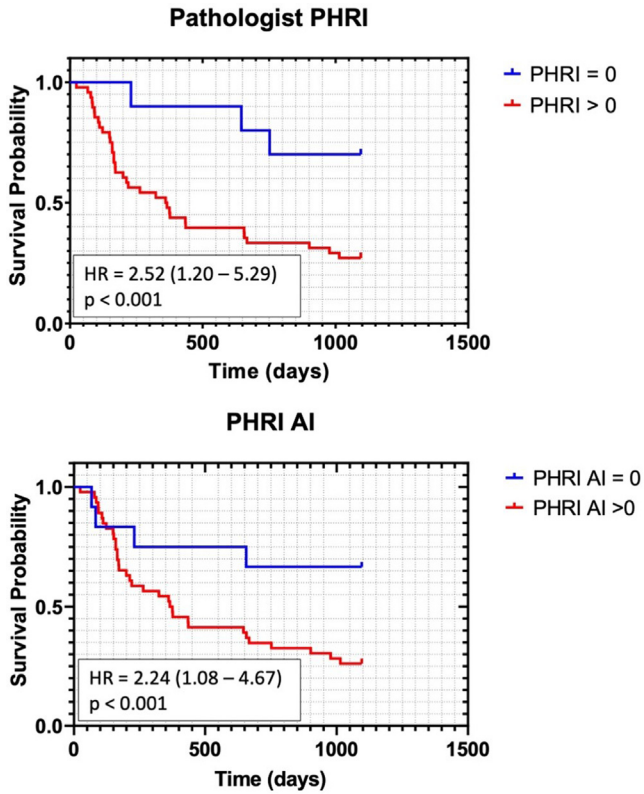or absence of UC activity in each WSI (biopsy). For this purpose, a weakly supervised learning, called MIL, was applied.

The deep learning model was trained end-to-end with a learning rate of 0.01 for 10 epochs. The loss function for the backbone with attention constraints was minimized with an L2 penalty and the UC activity prediction with the binary cross-entropy loss. The batch size was set equal to 1, so each WSI is analyzed in a separate way. This approach was implemented using Tensorflow 2.3.1 with Python 3.6. Experiments were conducted on the NVIDIA DGX A100 system.

To predict the occurrence of clinical outcomes, we used a contrastive learning paradigm based on the features extracted by the MIL model described previously. In particular, we used the embedding features obtained by the model as input. Afterward, we optimized a projection head that maps the WSI representation to a lower dimensionality space using the supervised contrastive loss. This function encourages the encoder to give closely aligned representations to entries from the same class, resulting in a more robust clustering of the representation space. Using many positive and negative pairs can improve the model's intra- and interclass variability. Afterward, the optimizer embedding was used for resolving survival prediction.

Supplementary Reference

1. Del Amor R, Launet L, Colomer A, et al. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in WSI. Artif Intell Med 2021; 121:102197.

Pathologist PHRI

PHRI AI

Supplementary Figure 1. The Kaplan Meier curves show the cumulative risk of suffering a disease flare (defined as any of the prespecified adverse clinical outcomes: UC-related surgery, UC-related hospitalization, UC-treatment dose optimization or medication initiation due to disease activity) after baseline endoscopy. The classification in histological remission or activity is based on PHRI (0 vs > 0) assessed by the human pathologist (*upper panel*) and the AI model (*lower panel*). The hazard ratios express the increase in risk of adverse events in the active group compared with the remission group.

Supplementary Table 1. CAD System Diagnostic Performance for the Prediction of Endoscopic Activity/Remission According to UCEIS and PICaSSO Scores

	CAD system diagnostic performance for prediction of endoscopic remission/activity	
	UCEIS (UCEIS >1)	PICaSSO (PICaSSO >3)
	Test (n = 373)	Test (n = 375)
Sensitivity	0.78 (0.70–0.85)	0.86 (0.77–0.92)
Specificity	0.80 (0.74–0.84)	0.78 (0.73–0.83)
PPV	0.65 (0.59–0.71)	0.60 (0.54–0.66)
NPV	0.88 (0.84–0.91)	0.93 (0.90–0.96)
F1 Score	0.71 (0.67–0.76)	0.71 (0.66–0.75)
Accuracy	0.79 (0.75–0.83)	0.80 (0.76–0.84)
AUROC	0.79 (0.75–0.83)	0.82 (0.78–0.86)