Journal Pre-proof

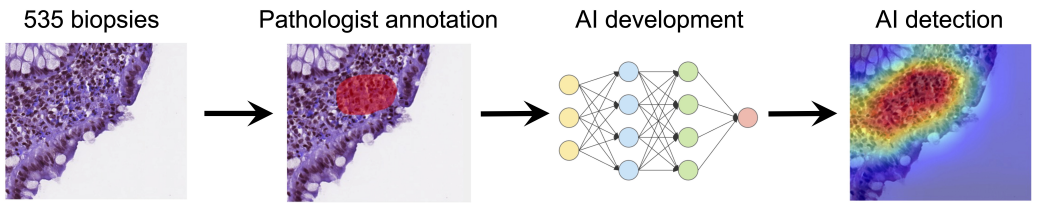
Artificial Intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis

Marietta Iacucci, Tommaso Lorenzo Parigi, Rocio del Amor, Pablo Meseguer, Giulio Mandelli, Anna Bozzola, Alina Bazarova, Pradeep Bhandari, Raf Bisschops, Silvio Danese, Gert De Hertogh, Jose G Ferraz, Martin Goetz, Enrico Grisan, Xianyong Gui, Bu Hayee, Ralf Kiesslich, Mark Lazarev, Remo Panaccione, Adolfo Parra-Blanco, Luca Pastorelli, Timo Rath, Elin S Røyset, Gian Eugenio Tontini, Michael Vieth, Davide Zardo, Subrata Ghosh, Valery Naranjo, Vincenzo Villanacci

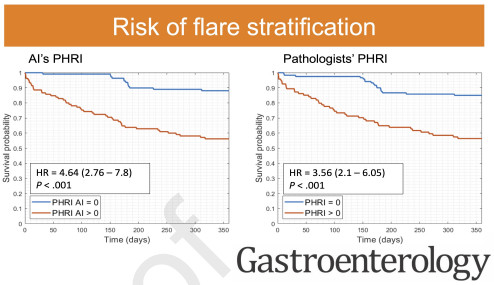
PII: S0016-5085(23)00216-0
DOI: https://doi.org/10.1053/j.gastro.2023.02.031
Reference: YGAST 65578

To appear in: *Gastroenterology*
Accepted Date: 15 February 2023

Please cite this article as: Iacucci M, Parigi TL, del Amor R, Meseguer P, Mandelli G, Bozzola A, Bazarova A, Bhandari P, Bisschops R, Danese S, De Hertogh G, Ferraz JG, Goetz M, Grisan E, Gui X, Hayee B, Kiesslich R, Lazarev M, Panaccione R, Parra-Blanco A, Pastorelli L, Rath T, Røyset ES, Tontini GE, Vieth M, Zardo D, Ghosh S, Naranjo V, Villanacci V, Artificial Intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis, *Gastroenterology* (2023), doi: https://doi.org/10.1053/j.gastro.2023.02.031.



AI assessment of histological remission/activity			
	PHRI >0	**Robarts > 3***	**Nancy > 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)



Gastroenterology

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], Vincenzo Villanacci[6]

*Authors share co-first authorship

1. Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK
2. NIHR Wellcome Trust Clinical Research Facility, University Hospital Birmingham, Birmingham, UK
3. Department of Gastroenterology, University Hospitals Birmingham NHS Trust, Birmingham, UK
4. APC Microbiome Ireland, College of Medicine and Health, University College 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, UK
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, 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, UK

17. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, USA
17. King's Health Partners Institute of Therapeutic Endoscopy, King's College Hospital, London, UK
18. Division of Gastroenterology, Helios HSK Wiesbaden, Wiesbaden, Germany
19. Department of Gastroenterology, Johns Hopkins Hospital, Baltimore, USA
20. NIHR Nottingham Biomedical Research Centre, Nottingham University Hospitals NHS Trust, Nottingham, UK
21. Department of Gastroenterology, University of Nottingham, Nottingham, UK
22. Department of Health Sciences, School of Medicine Ospedale San Paolo, Università degli Studi di Milano, Milan, Italy.
23. Department of Gastroenterology, Friedrich Alexander University of Erlangen, Nuremberg, Germany
24. Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway
25. Fondazione IRCCS Ca'Granda Ospedale Maggiore Policlinico, Milan, Italy
26. Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
27. Institute of Pathology, Friedrich-Alexander-University Erlangen-Nuremberg, Klinikum Bayreuth, Bayreuth, Germany
28. Department of Pathology, San Bortolo Hospital, Vicenza, Italy

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

MI: Study conception and design, Data acquisition, analysis and interpretation of data, drafting of the manuscript, Critical revision of the manuscript for important intellectual content.
TLP: Study conception and design, Analysis and interpretation of data, Drafting the manuscript, Critical revision of the manuscript for important intellectual content

RDA: Study conception, Analysis and interpretation of data, Statistical Analysis, Critical revision of the manuscript for important intellectual content
PM: Study conception, Analysis and interpretation of data, Statistical analysis, Critical revision of the manuscript for important intellectual content
GM: Data acquisition, Critical revision of the manuscript for important intellectual content
AB: Data acquisition, Critical revision of the manuscript for important intellectual content
XG: Data acquisition, Interpretation of data, critical revision of the manuscript for important intellectual content
AB: Statistical analysis, Critical revision of the manuscript for important intellectual content
PB: Data acquisition, Critical revision of the manuscript for important intellectual content
RB: Data acquisition, Critical revision of the manuscript for important intellectual content
SD: Critical revision of the manuscript for important intellectual content
GDH: Data acquisition, Critical revision of the manuscript for important academic content
JF: Data acquisition, Critical revision of the manuscript for important academic content
MG: Data acquisition, Critical revision of the manuscript for important academic content
EG: Analysis and interpretation of data, critical revision of the manuscript for important intellectual content
BH: Data acquisition, Critical revision of the manuscript for important intellectual content
RK: Data acquisition, Critical revision of the manuscript for important academic content
ML: Data acquisition, Critical revision of the manuscript for important academic content
RP: Data acquisition, Critical revision of the manuscript for important academic content
APB: Data acquisition, Critical revision of the manuscript for important academic content
LP: Data acquisition, Critical revision of the manuscript for important academic content
TR: Data acquisition, Critical revision of the manuscript for important academic content
ESR: Data acquisition, Critical revision of the manuscript for important intellectual content
GET: Data acquisition, Critical revision of the manuscript for important intellectual content
MV: Data acquisition, Critical revision of the manuscript for important intellectual content
DZ: Data acquisition, Critical revision of the manuscript for important intellectual content
SG: Study conception and design, analysis and interpretation of data, critical revision of the manuscript for important intellectual content
VN:Study conception and design, analysis and interpretation of data, Statistical Analysis, Critical revision of the manuscript for important intellectual content
VV: Study conception and design, Data acquisition, analysis and interpretation of data, critical revision of the manuscript for important intellectual content

Funding

The NIHR Birmingham Biomedical Research Centre funds MI 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

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.

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 diagnosis; AI, Artificial Intelligence; CNN, convolutional neural network;

Data Statement

The aggregate and anonymized data supporting the findings of this study are available within the article and its supplementary materials. The machine learning algorithm generated from the above-mentioned data is available from the corresponding author (MI) upon reasonable request.

Abstract

Background: 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: 535 digitalized biopsies (273 patients) were graded according to the PICaSSO Histologic Remission Index (PHRI), Robarts' (RHI), and Nancy Histological Index (NHI). 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% (RHI), and 89% and 79% (NHI). The model predicted the corresponding endoscopic remission/activity with 79% and 82% accuracy for UCEIS and PICaSSO, 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 histological remission/activity in biopsies of UC and predicts flare-ups. This can expedite, standardize and enhance histological assessment in practice and trials.

Keywords: *Ulcerative Colitis; Picasso Histologic Remission Index; Computer-aided diagnosis; Convolutional Neural Network; Robarts Histopathology index*

INTRODUCTION

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. Over 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 two AI models have been developed, the first, by Vande Casteele and colleagues, focused on the detection of eosinophils and their correlation with disease activity;[10] 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: 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 one 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 UK 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 UK and Brescia in Italy, with the same inclusion and exclusion criteria, who underwent colonoscopy with at least one 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).

Digital pathology

At least two 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 one biopsy. All samples were fixed in formalin, stained with hematoxylin and eosin, and digitally scanned at 40× (0.25µm per pixel) using Aperio Digital Pathology Scanning system (Leica Biosystem, Illinois, USA). The biopsies from the original cohort were assessed by one of the 6 expert IBD pathologists (GDH, XG, ESR, MV, VV, DZ) blinded to clinical and endoscopic information. In the external validation cohort biopsies were assessed by either VV or DZ. Histological activity was graded according to RHI, NHI and the newly developed PHRI. The cut-offs 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 three expert GI pathologists (VV, GM, AB). This was carried out by circling the areas with and without neutrophils using MicroDraw® (NAAT, France).

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 (WLE-HD) and scored using MES[14] and UCEIS[15], then in virtual chromoendoscopy (VCE; 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]

Computer-Aided Diagnosis 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 or WSI) were down-sampled to 20x resolution, divided into 512x512x3 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 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 appendix.

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 cut-offs was reported as sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F1-Score (F1S), accuracy (ACC) and Area Under the ROC curve (AUROC). AUROCs were compared with the DeLong test.

The 95% confidence intervals (CI) 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. 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). Data analysis was performed with Python (3.8) and Matlab (Mathworks Inc, US).

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 ⅔ and ¾ of biopsies were in histological 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 statistically similar for the three scores in a range between 0.77-0.85.

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 UC patients 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 positive predictive value (PPV) of 75% (95% CI 0.69 - 0.80), a negative predictive value (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 histological 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), a 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 cut-off 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 = 0.15; PHRI vs NHI p = 0.76).

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 two scores. Supplementary Table 1

CAD system prognosis prediction

Grouping patients in histological remission or activity based on pathologists' assessment, the hazard ratio between the two groups for suffering any pre-specified adverse clinical event, a proxy for flare-up, was 3.56 (95% CI 2.10-6.05) when classified according to PHRI, 4.28 (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 a 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 histological scores, estimate the corresponding endoscopic activity, and predict future clinical outcomes. Our model was developed on digitalized whole slide images 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. Histological 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), though, admittedly, the positive predictive values were more modest. Of course, the prime role of neutrophils in defining histological activity is also true in RHI and NHI.[12 13], which are endorsed for use in clinical trials by scientific societies, such as ECCO[20,21]. Our tool represents an important first step towards a greater expedition and standardization of histological 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 around 80% accuracy. Though 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 histological data, considered the most stringent assessment of UC activity, and observed a strong association between histological 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 histological 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. Firstly, 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 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, etc. 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 sub analysis based on biopsy quality.

As pointed out before, the variability among scores is a major obstacle to interpreting and comparing histological data. For this reason, we tested our system also on the two 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, histological 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, histological confirmation adds less prognostic information. Secondly, 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 only be used 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 three 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, histological reporting is still largely descriptive and non-standard, 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 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 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.

References

1. Ungaro R, Mehandru S, Allen PB, et al. Ulcerative colitis. The 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:S0168-8278(22)00024–1.

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. Inflammatory Bowel Diseases 2021:izab122.

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. Artificial Intelligence in Medicine 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.

Author names in bold designate shared co-first authorship

Figure 1. Artificial intelligence model

Caption Figure 1: Framework of the deep learning approach to detect UC activity. 1) Biopsies are digitalized into whole slide images (WSI). 2) A training set of WSIs are 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.

Figure 2. Examples of the CAD system's output

Caption Figure 2. Panels A, B and C: examples of biopsy annotations used to train the CAD system. Panels 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.

Figure 3. Kaplan Meier curves of clinical events in histological remission/activity groups

Caption Figure 3. 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 histological remission or activity was made by pathologist grading inflammation according to PHRI (panel A), RHI (panel B) and NHI (panel D). Panel C shows the same classification in PHRI activity/remission by the computer-aided diagnosis system. The hazard ratios express the increased risk of adverse events in the histologically active groups compared to the remission groups. Higher hazard ratios correspond to higher risk of event and better outcome stratification.

Table 1 Demographics characteristics

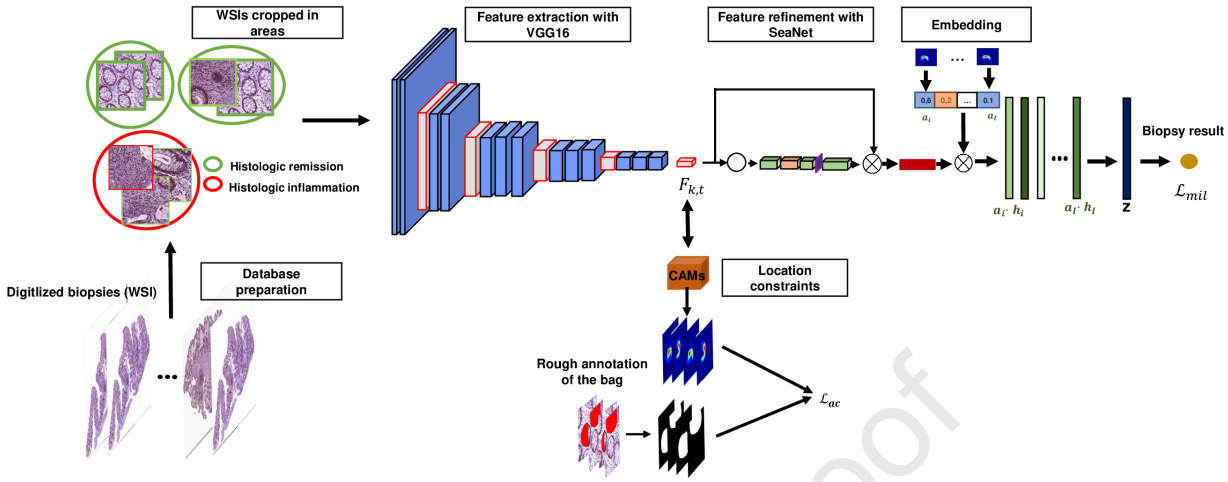
Characteristics		
Original Cohort		**Validation cohort**
Number of biopsies	535	154
Number of patients	273	58
Age mean ± sd	48.1 ± 14.8	44 ± 16
Female n (%)	111 (40.7)	25 (43)
Disease duration mean ± sd	14.6 ± 12.2	11.2 ± 9.3
Extension n (%)		
Left-sided colitis	116 (42.5)	23 (39.6)
Sub-total colitis 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*		

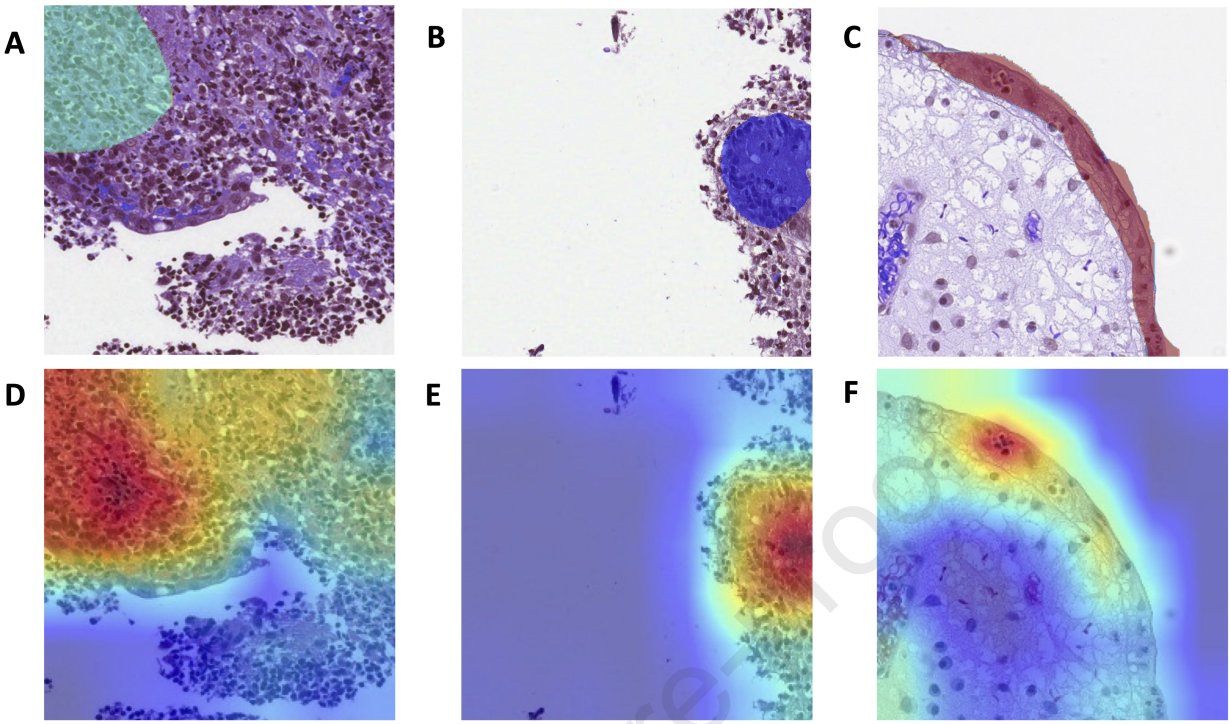
Remission (≤1) Active (>1) Missing data	371 (67.9) 172 (31.5) 3 (0.6)	48 (31.2) 103 (66.9) 3 (1.9)
PICaSSO*		
Remission (≤3) Active (>3) Missing data	418 (76.6) 126 (23.1) 2 (0.3)	Not available
Histology		
PHRI* Remission (=0) Activity (≥1) Missing Data	 342 (62.6) 200 (36.6) 4 (0.7)	 33 (21.4) 121 (78.6) 0 (0)
RHI* Remission (≤3 + no neutrophils) Activity (>3) Missing data	 413 (75.6) 128 (23.4) 5 (0.9)	 34 (22.1) 115 (74.7) 5 (3.2)
NHI* Remission (≤1) Activity (>1) Missing data	 389 (71.2) 152 (27.8) 5 (0.9)	 27 (17.5) 122 (79.2) 5 (3.2)

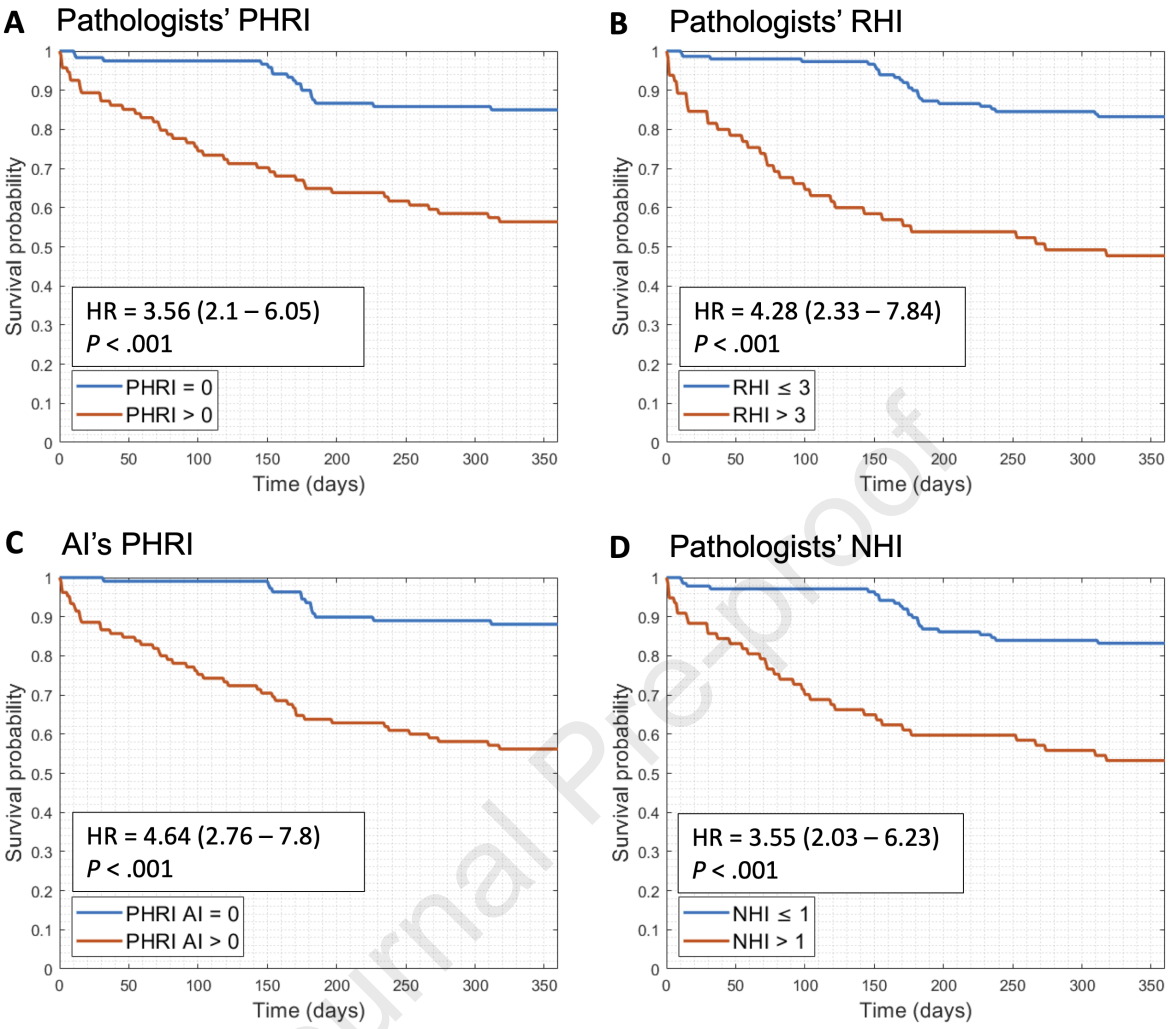
* UCEIS, PICaSSO, PHRI, RHI, NHI figures refer to biopsies (or site of biopsy), not patients
# and no neutrophils in the epithelium and the lamina propria

Table 2. CAD system diagnostic performance for detecting histological activity/ remission according to PHRI, RHI, NHI

CAD system diagnostic performance for histological remission/activity					
	PHRI (PHRI >0)		**RHI (RHI > 3)**	**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)







What do you need to know

Background and context: Histological remission is the optimal goal of treatment in ulcerative colitis, however histological assessment is limited by low agreement between pathologists.

New findings: Our validated AI model accurately distinguishes remission/inflammation according to 3 histological indices and stratifies risk of flare similarly to human physicians.

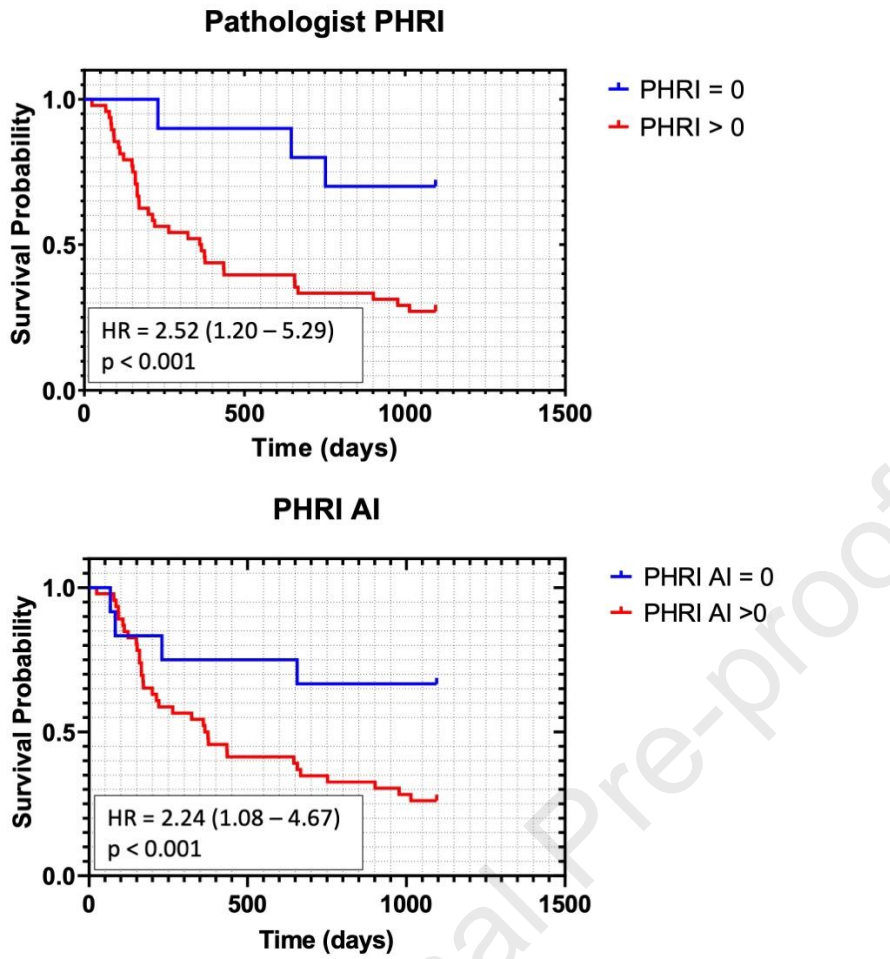
Limitations: The system cannot distinguish between different grades of disease severity.

Clinical research relevance: Our computer tool can speed up, simplify, and standardize histological 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.

Lay Summary

A newly developed artificial intelligence system was able to accurately distinguish remission from inflammation in biopsies of ulcerative colitis and predict prognosis.



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 to the remission group.

Supplemental material

Artificial Intelligence appendix

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 splitted the biopses collected in 11 centers). In a training subset of biopsies single neutrophils were annotated in the four regions of the biopsy (lamina propria, surface epithelium, cryptal epithelium and cryptal lumen) using an in-house software (MicroDraw). Whole-slide-images were then down-sampled to 20x resolution, divided into 512x512x3 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 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) [Rocío et al. AIIM 2021]. This model also focuses on neutrophils at patch level due to the attention constraints that obtains 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 multiple instance learning (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 a L2 penalty and the UC activity prediction with the binary cross-entropy loss. The batch size was set equal to one, 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 multiple instance learning model described above. In particular, we used the embedding features obtained by the model as input. Afterward, we optimised 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 inter-class variability. Afterward, the optimizer embedding was used for resolving survival prediction.

Reference: del Amor, R., Launet, L., Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., & Naranjo, V. (2021). An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in WSI. arXiv preprint arXiv:2104.09878.

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)

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	5
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	6-8
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	8-10
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	8
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	8
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	8
	5b	D;V	Describe eligibility criteria for participants.	8
	5c	D;V	Give details of treatments received, if relevant.	Na
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	10-11
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	9
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	9-11
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	9-11
Sample size	8	D;V	Explain how the study size was arrived at.	Na
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	8
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	Na
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	10-11
	10c	V	For validation, describe how the predictions were calculated.	10-11
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	11
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	Na
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	Na
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	9-11
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Na
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	11-12
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	12-13
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	12-13
	15b	D	Explain how to the use the prediction model.	13-15
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	12-13
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	Na
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	16
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	Na
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	16
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	15-17
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Na
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	3

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.