Detection, localisation, and quantification of neutrophils to assess disease activity and early response to therapy in ulcerative colitis: a novel AI-driven model

Marietta Iacucci,[a,*] Valentina Vadori,[b,i] Pablo Meseguer,[c,i] Irene Zammarchi,[a,i] Giovanni Santacroce,[a,i] Rocio Del Amor,[c,d,i] Davide Zardo,[e] Brian Hayes,[f,g] Rory Crotty,[f] Louise Burke,[f,g] Bisi Bode Kolawole,[b] Ujwala Chaudhuri,[b] Silvio Danese,[h] Subrata Ghosh,[a] Valery Naranjo,[c,d] and Enrico Grisan[b]

[a]APC Microbiome Ireland, College of Medicine and Health, University College Cork, Cork, Ireland
[b]School of Engineering, London South Bank University, London, UK
[c]Instituto Universitario de Investigación e Innovación en Tecnología Centarada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, València, Spain
[d]Artikode Intelligence S.L., Valencia, Spain
[e]Department of Pathology, San Bortolo Hospital, Vicenza, Italy
[f]Department of Histopathology, Cork University Hospital, Cork, Ireland
[g]Department of Pathology, University College Cork, Cork, Ireland
[h]Gastroenterology and Gastrointestinal Endoscopy Unit, Vita-Salute San Raffaele University and IRCCS San Raffaele Hospital, Milan, Italy

eClinicalMedicine
2025;90: 103658
Published Online xxx
https://doi.org/10.1016/j.eclinm.2025.103658

Summary

Background Histological remission (HR) is a key treatment goal in ulcerative colitis (UC), suggesting better disease management and treatment response. Absence of mucosal neutrophils is crucial to defining HR. However, the role of neutrophil quantity and localisation remains unclear. We aimed to develop a novel AI-driven pipeline to automate neutrophil detection, localisation, and quantification, supporting assessment of HR and treatment response.

Methods We developed an AI-driven pipeline by integrating and combining the outputs of two deep learning models to segment whole-slide images (WSIs) into epithelium, crypts and lamina propria and detect and quantify neutrophils. Optimal neutrophil density cut-offs to assess disease activity in UC patients from the AMAC phase 2 Mirikizumab trial (NCT02589665; 2015–2019) were identified and validated in the multicentre prospective PICaSSO cohort (2016–2019). Cut-offs to determine treatment response at weeks 12 and 52 were also evaluated.

Findings 303 WSIs of UC patients from the multicentre, randomised, double-blind, parallel-arm, placebo-controlled phase 2 clinical trial of Mirikizumab were analysed. The models yielded a 65.0% DICE Sørensen for region segmentation on average and 82.3% precision for neutrophil detection. A density cut-off $\geq$21.7 cells/mm^2 assessed disease activity with 86% (95% CI: 79%–92%) accuracy, 94% (88%–99%) sensitivity, 74% (54%–89%) specificity, as validated in the PICaSSO cohort. Moreover, a cut-off <21.7 cells/mm^2 demonstrated 79% (95% CI: 67%–92%) and 85% (95% CI: 70%–96%) accuracy in assessing treatment response at weeks 12 and 52, respectively.

Interpretation This novel AI-based pipeline demonstrates strong potential to detect, localise, and quantify neutrophils to assess histological activity in clinical trials and real-world settings. Furthermore, it effectively stratifies treatment response, offering a reliable and objective framework for personalised UC management.

Funding This paper has been supported by research grants from Eli Lilly and Company.

Copyright © 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Artificial intelligence; Neutrophils; Histological remission; Response to therapy; Ulcerative colitis

*Corresponding author. APC Microbiome Ireland, College of Medicine and Health, University College Cork, Cork, Ireland.
E-mail address: miacucci@ucc.ie (M. Iacucci).
[i]These authors contributed equally.

Research in context

Evidence before this study

Histological remission is a critical therapeutic target in ulcerative colitis. Nonetheless, its definition remains controversial. While the absence of neutrophils usually defines HR, their prognostic significance according to tissue localisation and quantity is still unclear. Moreover, their accurate evaluation is time consuming. Preliminary AI systems have been developed to evaluate the presence and quantity of neutrophils and predict clinical outcomes in UC. However, their role in assessing disease activity and response to therapy is still unexplored.

Added value of this study

We developed a novel AI-based system that accurately identifies, localises, and quantifies neutrophils in whole-slide images from UC patients in clinical trials and real-life settings. Moreover, it establishes neutrophil cut-offs to assess histological activity and therapeutic response.

Implications of all the available evidence

The AI pipeline offers a robust and objective alternative to manual histological scoring, addressing inter-rater variability and subjectivity. If further validated, this innovation has the potential to streamline clinical trial workflows and facilitate personalised therapeutic decisions based on precise neutrophil detection and quantification.

Introduction

Achieving histological remission (HR) has emerged as a critical treatment goal in ulcerative colitis (UC) being associated with reduced risk of relapse, hospitalisation, colectomy, and colorectal cancer.[1–4] Clinical trials are increasingly adopting more stringent endpoints, extending beyond endoscopic remission (ER) to target HR.[5,6] Nonetheless, the definition and consistent reproducible implementation of HR remain controversial.[7]

Several histological scoring systems have been developed to assess UC activity, emphasising different cellular features. Among these, the Geboes Score is widely used in clinical trials[8] but is limited by its complexity, lack of validation, and inclusion of numerous features. Its use is, therefore, still limited in clinical practice.[9] Recently, the European Crohn's and Colitis Organization (ECCO) underscored the pivotal role of neutrophils in determining UC activity, defining HR as their complete absence from both the epithelium and lamina propria.[10] Building on this definition, the validated and simple PICaSSO Histological Remission Index (PHRI) streamlined HR assessment by focussing solely on neutrophils in lamina propria or epithelium and crypts. This score strongly correlated with other histologic scores and effectively predicted long-term clinical outcomes.[11]

While neutrophils are well recognised as drivers of UC activity, their prognostic significance based on tissue localisation, epithelium versus lamina propria, requires further exploration. Evidence suggests that epithelial neutrophils, compared to lamina propria involvement, are more strongly associated with adverse outcomes in UC.[12] Additionally, quantifying neutrophils and identifying optimal cut-offs indicative of disease activity could advance both clinical trial workflows and disease management in clinical practice.

However, histological assessment by pathologists is time-consuming, requires specialised expertise, and is prone to significant intra- and inter-observer variability, even among experienced pathologists. These challenges limit the broader application of histological assessment in clinical trials and routine practice.[13] Artificial intelligence (AI) offers a promising solution to overcome these barriers. AI models have been successfully employed to assess neutrophils in UC whole-slide images (WSIs).[14,15] Notably, our PHRI-based AI model demonstrated high accuracy in assessing disease activity and predicting outcomes.[14] Similarly, a preliminary study highlighted the correlation between AI-detected neutrophil quantity and clinical outcomes.[16] Beyond accurate detection, AI can be pivotal in identifying neutrophil density cut-offs to assess disease activity and therapy response, addressing a significant unmet need in clinical trials and practice.

In this research, we developed a novel AI-driven pipeline to automate neutrophil detection, localisation, and quantification. This framework provides a standardised approach to evaluate histological activity, leveraging data from a global phase two clinical trial in UC. Its performance was validated in a real-world cohort of UC patients from the prospective multicentre international PICaSSO (Paddington International Virtual Chromoendoscopy Score) study.[17] Furthermore, we sought to identify optimal neutrophil cut-offs to facilitate automated assessment of early response to therapy in clinical trials, aiming to enhance trial efficiency and personalise UC management.

Methods

Patient characteristics

AI development: Mirikizumab phase 2 trial cohort

Data were collected from adult patients (18–75 years old) with moderate-to-severe active UC, defined as total

Mayo score of 6–12 and Mayo Endoscopic Subscore (MES) ≥2, participating in a multicentre, randomised, double-blind, parallel-arm, placebo-controlled phase 2 clinical trial evaluating the efficacy and safety of Mirikizumab in UC (NCT02589665).[18]

Baseline demographic and clinical features were recorded. Clinical remission was defined as Mayo scores of 0 for rectal bleeding, 0 or 1 (with a 1-point decrease from baseline) for stool frequency, and 0 or 1 for centrally-read endoscopy at W12 and 52.

AI validation: PICaSSO cohort

External validation was conducted in adult patients with UC undergoing endoscopy for disease assessment or surveillance from the prospective multicentre observational PICaSSO study (approved by the West Midlands Research Ethics Committee, 17/WM/0223).[17] Baseline demographic and clinical data were collected.

Endoscopic assessment

In the development cohort, endoscopic activity was assessed at baseline (before starting Mirikizumab), after induction (W12), and after maintenance (W52). High-definition white-light endoscopy videos were collected at different time points. ER was defined as MES 0 by a single central reader pool of three central readers.

In the validation cohort, disease activity was assessed through colonoscopy/sigmoidoscopy performed by a pool of twelve experienced endoscopists, and ER was defined as MES 0.[17]

Histologic assessment

Biopsies were taken from the most affected area, at least 30 cm from the anal verge. All samples were fixed in formalin, stained with haematoxylin and eosin, and digitally scanned at ×40 (0.25 µm per pixel). Digitisation scanners were similar between the two cohorts.

In the development cohort, histologic activity at each timepoint was scored by a central reader pool of two expert pathologists reviewing WSIs. HR was defined as Geboes ≤2B.0.

In the validation cohort, disease activity was assessed by six experienced pathologists using Geboes and PHRI. HR was defined as Geboes ≤2B.0 or PHRI = 0 (absence of neutrophils in epithelium and lamina propria). In the PICaSSO dataset, approximately 30% of patients had differing scores between the sigmoid and rectum. For the evaluation of the pipeline, the histological samples were assessed independently. Details on histological slides preprocessing and scoring systems are provided in Supplementary Material and Supplementary Table S1.

Response to therapy at weeks 12 and 52

The response to Mirikizumab at weeks 12 and 52 was assessed based on the achievement of histologic improvement (Geboes <3.1 or PHRI ≤1) or HR (Geboes ≤2B.0 or PHRI 0).

Development of a novel AI pipeline for neutrophil localisation, density quantification and cut-off calculation

An AI-based pipeline integrated two deep learning models to (1) segment WSIs into regions of surface epithelium, crypts and lamina propria and (2) detect the density of neutrophils. The outputs from these models were combined to calculate neutrophil densities (cells/mm^2) in the lamina propria, surface epithelium, crypts and overall tissue and establish optimal cut-offs for assessment of inflammation and response to therapy. The AI pipeline is schematically represented in Supplementary Figures S1 and S2.

Initially, the region segmentation model obtains the slide-level semantic segmentation of the slide to demarcate the lamina propria, surface epithelium and crypts of tissue section. The cell segmentation model[18] was used to identify the neutrophils. The region segmentation model was implemented using a Feature Pyramid Network (FPN),[19] while neutrophil segmentation model was performed with CISCA-Net, a lightweight U-Net–based model with three output heads for cell instance segmentation and classification. Additional implementation details are provided in the Supplementary Materials and in the original FPN and CISCA-Net[18] papers.

To classify the detected neutrophils as either "epithelial" or "lamina propria", the outputs of the cell and region segmentation models were merged. Each segmented cell was precisely mapped to a region based on the location of the cell centroid. Neutrophil density was then calculated by dividing the number of cells in each region by the region's area.

Optimal cell density cut-offs were determined to assess histologic activity and early therapeutic response at weeks 12 and 52. Cut-offs were estimated by maximising sensitivity (for disease activity assessment) or specificity (for improvement/remission) while ensuring that specificity or sensitivity remained higher or equal to 70.0%. The optimal cut-off was computed using the cutpoints R package[20] for each cell density by averaging the optimal cut-offs across 50 bootstrap samples.

Details on pipeline development are provided in Supplementary Material.

The pipeline's performance in assessing histological activity was evaluated using WSIs from the Mirikizumab phase 2 trial and then validated in the PICaSSO cohort. The AI's capability to identify the cut-off of neutrophil density to assess response to therapy based on histological improvement and remission was assessed at W12 and W52 in the Mirikizumab cohort.

Ethics

The AMAC Phase 2 Mirikizumab trial was approved by the Institutional Review Boards or Independent Ethics Committees at each participating centre. This study was compliant with the International Committee for Harmonisation guideline on good clinical practice. All informed consent forms and protocols were approved by appropriate ethical review boards before initiation of the study. All patients gave written informed consent before receiving the study drug.

The PICaSSO study was approved by the research ethics committee (17/WM/0223) for the UK centres. All international sites gained local research ethics committee approvals in their respective countries. All patients provided written informed consent for study participation, data collection, and sample analysis.

Statistics

The diagnostic performance of the cut-offs was analysed using the following metrics: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Area Under the Curve (AUC). Results are based on the Mirikizumab phase 2 trial data. Out-of-sample validation performance and 95% confidence intervals (CIs) for each metric were assessed using out-of-bag estimates derived from 1000 bootstrapped samples of the Lilly data. The PICaSSO data served as an external validation dataset for disease assessment.

Pipeline development was conducted in line with TRIPOD recommendations.

The correlation between the density of neutrophils (overall and in epithelium/lamina propria) and Geboes or PHRI score was assessed through the Spearman coefficient. Spearman correlation was selected over Pearson because visual inspection indicated a monotonic but non-linear relationship between histological scores and AI-detected neutrophil counts; thus, a rank-based method was deemed more appropriate. Correlations were interpreted as strong (≥0.7), moderate (0.4–0.7), weak (0.2–0.4), or absent (<0.2).[21]

Agreement between cut-off–based classifications and pathologist-assessed disease activity was evaluated through Cohen's kappa, interpreted according to the Landis and Koch framework[21] (<0.00: poor; 0.00–0.20: slight; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: substantial; 0.81–1.00: almost perfect agreement). Sample size and power analyses were conducted as part of the clinical trial design. In this study, however, the focus was on the development and validation of the computational pipeline for automated histological assessments rather than on trial-level powering. Variability in cut-off performance is reported with confidence intervals.

Data analysis was performed with R v4.0 (R Foundation for Statistical Computing, Vienna).

Details about the statistical plan can be found in Supplementary Material.

Role of funding source

This study was supported by Eli Lilly and Company. The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Population characteristics and disease-related outcomes

Demographic and clinical characteristics of the study population and disease-related outcomes at different time points are summarised in Table 1. No baseline demographic data were missing among the patients considered in the cohorts.

AI development: Mirikizumab phase 2 clinical trial cohort

A total of 106 patients with moderate-to-severe UC from the phase 2 Mirikizumab clinical trial were included. 46.2% (49/106) were females, and average age was 37.5 years (29.0–53.3). 17.9% (19/106) of patients achieved clinical remission after 12 weeks and 41.3% (38/92) after 52 weeks of treatment. 5.7% (6/106) of patients achieved ER after 12 weeks and 20.6% (19/92) after 52 weeks of treatment in the Mirikizumab cohort. 60.3% (64/106) and 47.2% (50/106) of patients exhibited histologic improvement and remission at W12, respectively. 69.2% (63/91) and 57.1% (52/91) of patients achieved histologic improvement and remission at W52, respectively.

AI validation: PICaSSO cohort

302 patients from the prospective multicentre international PICaSSO study were considered. 40.7% (125) were females; the average age was 48.4 ± 14.8. 76.2% (234/302) of patients were on treatment with 5-aminosalicylic acid, while 39.1% (118/302) were on biological therapy. 65.2% (197/302) were in clinical remission, 55.6% (168/302) in ER and 59.6% (180/302) in HR during disease assessment in the PICaSSO cohort.

AI pipeline diagnostic performance

303 WSIs from active UC patients in the Mirikizumab cohort were analysed. WSIs from 15 patients at week 52 were missing due to patient dropout. No exclusion criteria based on image quality were applied.

The AI region segmentation model yielded DICE–Sørensen coefficients of 85.2% for lamina propria, 44.8% for surface epithelium, and 64.9% for crypts. The AI model for neutrophil detection demonstrated strong overall performance, with 82.3% (±15.8%) precision, 77.0% (±17.8%) recall, and 77.5% (±12.1%) F1 score.

Correlation between the density of neutrophils and Geboes and PHRI histological score

A strong positive correlation was found between neutrophil density and grades of Geboes score at all time points [0.81 (0.77–0.85)], with strong correlation at W12 [0.79 (0.70–0.85)] and W52 [0.78 (0.68–0.85)]. Similarly, a strong positive correlation was found between neutrophils density and PHRI score [0.70 (0.65–0.74)] in the PICaSSO cohort. Details on the correlation between neutrophil density, location and Geboes grades are presented in Fig. 1.

AI performance to quantify neutrophils cut-off for disease assessment based on Geboes and PHRI

Overall, 192 out of 303 biopsies in the Mirikizumab development cohort showed histological activity. In comparison, 119 out of 438 biopsies showed histological activity in the PICaSSO validation cohort.

The pipeline identified an overall optimal neutrophil density cut-off of $\geq$21.7 cells/mm^2 to assess disease activity according to Geboes and PHRI, with 86% (79%–92%) accuracy, 94% (88%–99%) sensitivity, 74% (54%–89%) specificity and 95% (92%–98%) AUC.

In the validation cohort, the same cut-off showed robust performance for disease activity assessment by Geboes and PHRI, with 87% (84%–90%) accuracy, 69% (61%–77%) sensitivity and 94% (92%–97%) specificity. Similarly, 85% (82%–89%) accuracy, 62% (54%–70%) sensitivity and 97% (94%–98%) specificity were found for PHRI.

When considering the optimal cut-off to assess disease activity according to neutrophil localisation, a region-specific cut-off of 30.4 cells/mm^2 was found for the lamina propria, which showed an 85% (79%–91%) accuracy, 93% (87%–98%) sensitivity and 72% (55%–90%) specificity. A cut-off of 10.1 cells/mm^2 in the epithelium showed an 84.9% (78%–91%) accuracy, 90% (95% CI 83%–96%) sensitivity and 75% (56%–94%) specificity for assessment of disease activity.

Table 2 and Supplementary Table S2 provide detailed results on neutrophil cut-offs to assess disease activity. Fig. 2 reports the graphic distribution of neutrophil density in both cohorts. Supplementary Table S3 provides a subanalysis of the pipeline performance focussing on WSI from W0.

		Mirikizumab cohort (N = 106)	PICaSSO cohort (N = 302)
Age (years), median (IQR)		37.5 (29.0–53.3)	48.4 (33.6–63.2)
Sex (female), n (%)		49 (48.0)	125 (40.7)
Disease length (years), median (IQR)		5.52 (2.9–10.2)	15.0 (4.2–25.8)
Prior biologic exposure, n (%)		50 (49.0)	118 (38.4)
Corticosteroids use, n (%)		46 (45.1)	74 (24.1)
Amino salicylates use, n (%)		82 (80.4)	234 (76.2)
Thiopurines use, n (%)		27 (26.5)	68 (22.1)
Clinical remission, n (%)	W12	19 (17.9)	197 (65.2)
	W52	38 (41.3)	
Endoscopic remission, n (%)	W0	0 (0.0)	168 (55.6)
	W12	6 (5.7)	
	W52	19 (20.6)	
Histological remission, n (%)	W0	9 (8.5)	180 (59.6)
	W12	50 (47.2)	
	W52	52 (57.1)	

Abbreviation: IQR: interquartile range.

Table 1: **Patients characteristics.**

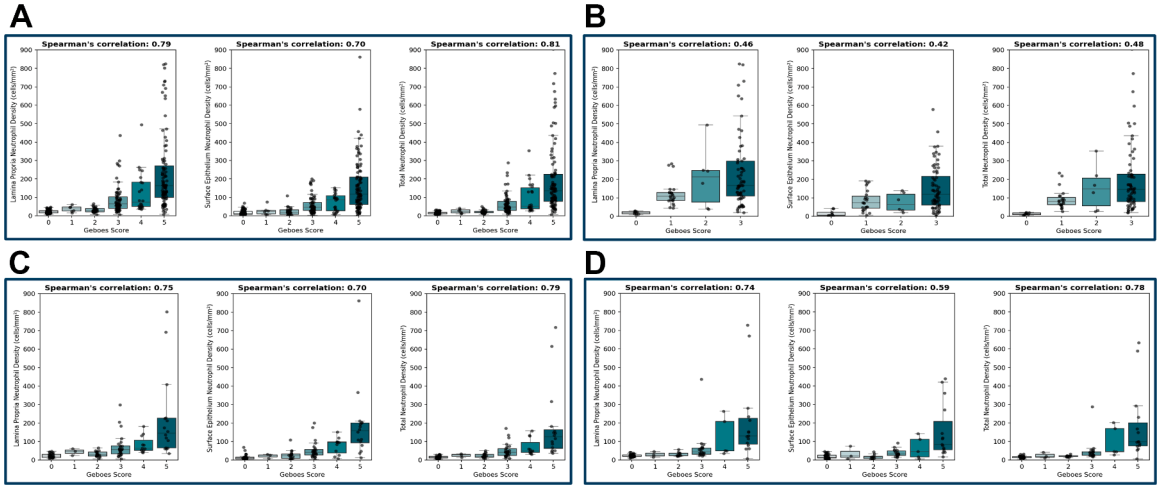

Fig. 1: **Spearman correlation between neutrophil density and subcategories of the Geboes score.** (A) Overall correlation across all time points; (B) baseline (week 0); (C) post-induction (week 12); and (D) post-maintenance (week 52). Each dot represents a whole-slide image (WSI). Correlation coefficients (ρ) were computed using the Spearman rank test. Error bars represent 95% confidence intervals (CIs) for correlation coefficients.

Assessment of disease activity (n = 303, n_pos = 192, n_neg = 111)—Geboes and PHRI

Neutrophil localisation	Cut-off (cells/mm^2)	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score	AUC
Surface epithelium	24.9	79% (72%–85%)	83% (76%–90%)	71% (53%–87%)	84% (74%–92%)	71% (60%–82%)	83% (78%–88%)	88% (83%–93%)
Surface + cryptal epithelium	10.1	85% (78%–91%)	90% (83%–96%)	75% (56%–94%)	86% (77%–97%)	84% (71%–91%)	88% (83%–93%)	93% (89%–97%)
Lamina propria	30.4	85% (79%–91%)	93% (87%–98%)	72% (55%–90%)	86% (77%–94%)	86% (77%–96%)	89% (84%–94%)	94% (91%–97%)
Total	21.7	86% (79%–92%)	94% (88%–99%)	74% (54%–89%)	86% (76%–94%)	87% (78%–96%)	90% (84%–94%)	95% (92%–98%)

Abbreviations: AUC: area under the curve; NPV: negative predictive value; PHRI: PICaSSO Histological Remission Index; PPV: positive predictive value.

Table 2: **Cut-off of neutrophils to assess disease activity by Geboes and PHRI–Mirikizumab cohort.**

Neutrophil cut-off to assess Mirikizumab phase 2 clinical trial response to treatment at W12 and W52

Neutrophil cut-off to assess Mirikizumab phase 2 clinical trial histological improvement at W12 and W52 based on Geboes and PHRI. 106 and 91 patients were considered to assess histological improvement at W12 and W52, respectively. 64 (60.4%) out of 106 and 63 (69.2%) out of 91 achieved histological improvement at W12 and W52 by Geboes. 60 (56.6%) out of 106 and 53 (58.2%) out of 91 achieved histological improvement at W12 and W52 by PHRI. Shared cut-offs were calculated for both weeks.

An overall neutrophil cut-off of 25.0 and 22.9 cells/mm^2 stratified histological improvement at W12 and W52 according to Geboes and PHRI, respectively. At W12, these cut-offs corresponded to accuracies of 82% (67%–93%) and 81% (68%–93%), sensitivities of 75% (50.0%–96%) and 75% (50.0%–96%), specificities of 93% (80%–100%) and 90% (77%–100%), and AUCs of 93% (87%–97%) and 91% (84%–97%). At W52, the same cut-offs corresponded to accuracies of 84% (68%–95%) and 84% (69%–96%), sensitivities of 78% (56%–96%) and 80% (53%–100%), specificities of 96% (83%–100%) and 90% (75%–100%), and AUCs of 92% (80%–97%) and 92% (82%–98%), for Geboes and PHRI, respectively.

The optimal neutrophil density cut-off to assess histological improvement was higher in the lamina propria than epithelial compartment (33.7 vs. 11.8 cells/mm^2 for Geboes and 32.2 vs. 11.2 cells/mm^2 for PHRI). Further details about the pipeline performance for epithelium and lamina propria are detailed in Supplementary Table S4 and Supplementary Table S5.

Neutrophil cut-off to assess HR at W12 and W52 based on Geboes and PHRI. 50 (47.2%) out of 106 and 52 (57.1%) out of 91 achieved HR at W12 and W52.

A neutrophil cut-off of 21.7 cells/mm^2 effectively stratified histological remission at both W12 and W52, yielding accuracies of 79% (67%–92%) and 85% (70%–96%), sensitivities of 74% (45%–96%) and 80% (53%–100%), specificities of 84% (71%–96%) and 90% (77%–100%), and AUCs of 91% (83%–97%) and 93% (84%–98%) at W12 and W52, respectively. The optimal neutrophil density cut-off to assess HR was higher in the lamina propria than in the epithelium compartment (29.6 vs. 10.6 and cells/mm^2).

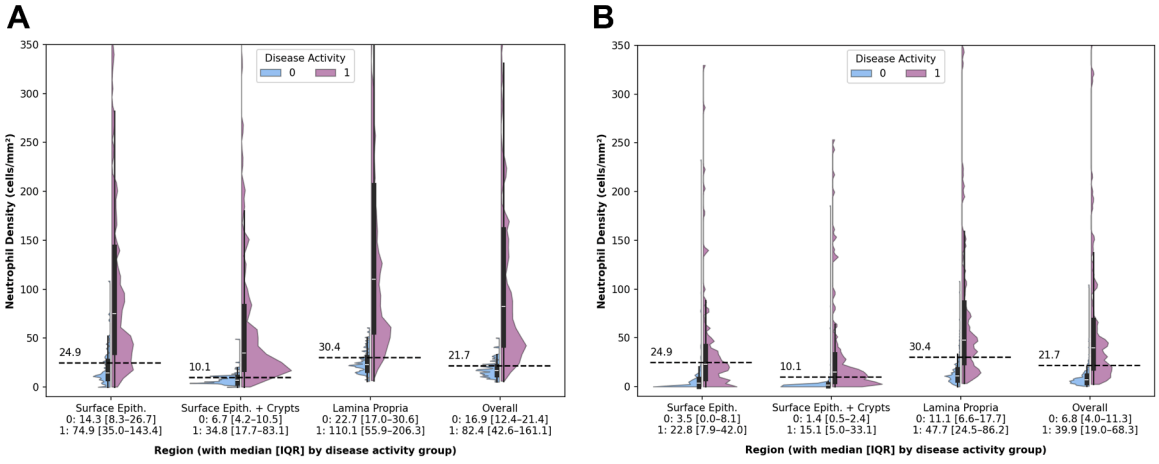

Fig. 2: **Neutrophil density distribution in the Mirikizumab and PICaSSO cohorts.** (A) Distribution of neutrophil density (cells/mm^2) across all tissue compartments in the Mirikizumab cohort. (B) Distribution in the PICaSSO validation cohort according to localisation: surface epithelium, crypts, and lamina propria. Dashed lines indicate optimal AI-derived cut-offs for distinguishing histological activity and remission. Box plots display median values (horizontal line), interquartile range (IQR; box), and whiskers representing 1.5 × IQR. p-values were calculated using the Mann–Whitney U test. Abbreviations: epith: epithelium; IQR: interquartile range.

Table 3 provides further details about the pipeline performance for epithelium and lamina propria.

Agreement between cut-off–based classifications and disease activity

Overall agreement between cut-off–based classifications and disease activity as assessed by pathologists in the Mirikizumab cohort was substantial (κ = 0.72). Agreement was moderate for surface epithelium (κ = 0.55) and substantial for surface epithelium + crypts (κ = 0.64) and lamina propria (κ = 0.70). A similar trend was observed in the PICaSSO cohort.

At week 12, agreement between cut-off–based classifications and remission was moderate across regions (κ = 0.47–0.54), but improved by weeks 52, ranging from moderate for surface epithelium (κ = 0.50) to substantial for surface epithelium + crypts (κ = 0.63), lamina propria (κ = 0.63), and overall (κ = 0.72).

Overall agreement between cut-off–based classifications and histologic improvement at week 12 was moderate (Geboes κ = 0.58; for surface epithelium κ = 0.54 and lamina propria κ = 0.58). It was substantial for surface epithelium + crypts (κ = 0.70). By week 52, agreement was moderate for surface epithelium (κ = 0.49) and surface epithelium + crypts (κ = 0.59), but reached substantial levels for lamina propria (κ = 0.66) and overall (κ = 0.68).

Further details on the agreement across cohorts and different time points are provided in Supplementary Table S6.

Discussion

To the best of our knowledge, this study developed the first AI pipeline to accurately detect, localise and quantify neutrophils in WSIs from a clinical trial cohort to assess disease activity and response to therapy. This pipeline identified neutrophil density cut-offs with remarkable ability, demonstrating 86% (79%–92%) accuracy, 94% (88%–99%) sensitivity, 74% (54%–89%) specificity in assessing disease activity, confirmed in a real-world validation cohort. Additionally, the pipeline demonstrated novel utility in identifying cut-offs that assess response to Mirikizumab at different time points. Notably, differences in neutrophil localisation within epithelium, crypts and lamina propria revealed varying predictive abilities, with lower epithelial and cryptal neutrophil counts correlating with disease remission and therapy response.

This automated pipeline marks a significant advancement in addressing unmet needs in clinical trials. HR has emerged as a critical therapeutic target in UC, being adopted in combination with ER as a primary endpoint in several trials.[10,22] However, HR remains inconsistently defined, and multiple scoring systems currently used are hindered by complexity, subjectivity, and inter-/intra-rater variability.[8,11,23,24]

Assessment of response to therapy—Histological remission (W12: n = 106, n_pos = 50, n_neg = 56; W52: n = 91, n_pos = 52, n_neg = 39)—Geboes and PHRI

	Cut-off (cells/mm^2)	Accuracy		Sensitivity		Specificity		PPV		NPV		F1 score		AUC	
		W12	W52	W12	W52	W12	W52	W12	W52	W12	W52	W12	W52	W12	W52
Surface epithelium	24.4	74% (63%–87%)	75% (59%–88%)	73% (44%–100%)	74% (44%–94%)	76% (59%–93%)	77% (55%–94%)	73% (57%–89%)	81% (65%–95%)	77% (59%–100%)	70% (47%–92%)	72% (53%–88%)	77% (56%–90%)	84% (74%–94%)	85% (74%–94%)
Surface + cryptal epithelium	10.6	82% (67%–92%)	82% (69%–94%)	87% (50%–100%)	76% (55%–96%)	77% (59%–93%)	90% (75%–100%)	77% (61%–92%)	91% (78%–100%)	89% (66%–100%)	74% (54%–94%)	81% (57%–93%)	82% (67%–95%)	87% (77%–96%)	91% (82%–98%)
Lamina propria	29.6	78% (75%–89%)	81% (69%–92%)	74% (44%–95%)	77% (53%–95%)	83% (67%–96%)	87% (69%–100%)	79% (63%–94%)	89% (75%–100%)	79% (62%–95%)	74% (57%–93%)	76% (56%–88%)	82% (67%–93%)	89% (81%–96%)	91% (81%–97%)
Total	21.7	79% (67%–92%)	85% (70%–96%)	74% (45%–96%)	80% (53%–100%)	84% (71%–96%)	91% (77%–100%)	80% (64%–95%)	93% (81%–100%)	79% (61%–96%)	79% (57%–100%)	76% (57%–91%)	85% (67%–97%)	91% (83%–97%)	93% (84%–98%)

Abbreviations: AUC: area under the curve; NPV: negative predictive value; PHRI: PICaSSO Histological Remission Index; PPV: positive predictive value. No difference between the pathologists' evaluation of remission based on Geboes or PHRI was seen, meaning that the cut-offs are identical for assessment of disease activity assessment and assessment of remission at week 12/52 between the two scores.

Table 3: **Cut-off of neutrophils to assess histological remission at W12 and W52 based on Geboes and PHRI.**

Although consensus broadly agrees that neutrophils are key indicators of UC activity, rare or scattered neutrophils, especially in the lamina propria, may still be consistent with HR.[10] Therefore, the potential impact of such small number of neutrophils, often difficult to detect and quantify by human assessment, on adverse outcomes requires further investigation. Moreover, manual assessment by pathologists is time-consuming and prone to bias.

In our cohort, rates of HR were higher than ER. This discrepancy may reflect differences between histology and endoscopic activity and sampling variability from non-targeted biopsies, or biased assessment of histological activity. Our automated pipeline accurately identified and quantified neutrophils, facilitating assessment of disease activity in a clinical trial cohort. This approach may complement human interpretation and define histological healing through AI-quantified small numbers of neutrophils in epithelium and lamina propria.

Its strong correlation with pathologist-assessed Geboes and PHRI scores at weeks 12 and 52 reinforced its reliability.[8,11] Importantly, there was also moderate-to-substantial agreement between cut-off–based classification and disease activity as assessed by pathologists across different compartments (overall, epithelium, surface epithelium + crypts, lamina propria), further proving the ability of the model to provide expert-like disease assessment. Traditional histological scores rely on semi-quantitative grading, which is inherently subjective and prone to inter-/intra-observer variability. In contrast, AI derived cut-offs offer an objective, reproducible, and fully quantitative approach. Moreover, the compartment-specific thresholds identified by the pipeline provide granular insights into disease activity not captured by global scoring systems. This suggests that AI quantification could complement existing scores by detecting subtle neutrophil distributions that are difficult for pathologists to consistently recognise, thereby improving precision in predicting therapeutic response and outcomes.

Initially, pipeline development focused on a cohort of active patients, ensuring high presence of neutrophils to build an accurate detection and quantification framework, determining precise cut-off of neutrophils.

To evaluate the robustness of our AI pipeline, it was also tested in the external real-world international PICaSSO cohort, where clinical, endoscopic, and HR is predominant. This allowed us to confirm the pipeline's reliability and robustness to effectively detect few numbers and low density of neutrophils in both lamina propria and epithelium. Despite showing slightly lower sensitivity than the development cohort, the original cut-offs maintained high performance for assessing disease activity (87% [84%–90%] accuracy, 69% [61%–77%] sensitivity, 94% [92%–97%] specificity), confirming its reproducibility in clinical practice.

Preliminary evidence suggests that automated neutrophil cut-offs could predict major adverse outcomes in real-world settings,[16] being more predictive of therapeutic response and early flare-ups. Our pipeline can uniquely automate, quantify, and localise neutrophils within the epithelium, crypts and lamina propria, potentially complementing and surpass the qualitative evaluation of most scores. While slight variations in neutrophils may not influence treatment decisions, a broader understanding of neutrophil distribution and quantification might be crucial in assessing treatment responsiveness and guiding therapeutic strategies.

Emerging evidence suggests that epithelial neutrophils are more predictive of clinical outcomes than their lamina propria counterparts.[12] Region-specific cut-offs predictive of HR were identified, with lower thresholds for the epithelium and crypts than lamina propria. These findings were again validated in the PICaSSO cohort. Epithelial, cryptal and lamina propria neutrophil density consistently demonstrated stronger predictive value for therapy response, reflecting their role in acute mucosal injury. These observations indicate that epithelial neutrophil density may serve as a more sensitive biomarker of disease activity or therapeutic efficacy, whereas lamina propria infiltration reflects longer-term inflammatory tone.

The AI multi-model pipeline also demonstrated a strong ability to assess therapeutic response at W12 and W52 in the trial development cohort. This ability addresses a significant gap in both clinical trials and practice. By refining patient stratification, this pipeline could improve optimisation of induction and maintenance treatments, maximising therapeutic efficacy while minimising side effects and costs. In clinical practice, it can guide physicians in making informed therapeutic decisions, ultimately improving patient management and outcomes.

Interestingly, epithelial neutrophils demonstrated a lower cut-off for assessing therapy response, highlighting their pivotal role in disease progression and treatment outcomes. In addition, the low cut-off for epithelial and cryptal epithelium to evaluate disease activity and response to therapy highlights the importance of cryptal infiltration as a hallmark of severe disease. The ability of the pipeline to reliably quantify cryptal neutrophils, despite segmentation challenges, provides a robust tool for refining disease stratification.

Several key innovations distinguish our research from previous works. Firstly, our pipeline employs a refined multi-scale architecture specifically trained to detect, segment and quantify neutrophils in clinical trial and reproduced in real-world cohort. Second, it defines a neutrophil cut-off for both inflammation assessment and treatment response, addressing an unmet need in clinical trials. Third, the pipeline was validated across an independent, multicentre external dataset (PICaSSO cohort), crucial for demonstrating

reproducibility in real-world setting where consistency is essential. By standardising histological assessment, our pipeline offers a transformative tool for clinical trials, reducing inter-rater variability, streamlining workflows, and lowering operational costs associated with central readout. Integration into existing workflows could enable AI-based quantification to replace manual neutrophil scoring for trial endpoints, while pathologists remain involved in the evaluation in complex or discordant cases. Compared to manual scoring, which requires several minutes per slide, the AI pipeline delivers fully automated results within seconds, reducing both turnaround time and labour-associated costs.

The research also has limitations. The pipeline focuses exclusively on neutrophils, given their primary role in determining disease activity and outcome prediction compared to other cells,[10] such as eosinophils and plasma cells.[11] Future work will expand the pipeline to multi-feature analysis for more comprehensive activity assessment. Furthermore, the dataset is based on the specific cohort of phase 2 trial on Mirikizumab, which may limit broader applicability. Other datasets and cohorts evaluating patients on other advanced therapies should be analysed to validate the model and assess the response to treatment with different mechanisms of action.

Variations in staining protocols and slide preparation across cohorts may have impacted the pipeline's performance. Nonetheless, they did not lead to significant inconsistencies in results.

Segmentation accuracy was generally high, and the DICE coefficient for epithelium was relatively low and moderate for crypts. Surface epithelium and crypts segmentation remains challenging due to their biological variability and susceptibility to artefacts, particularly in inflamed tissue. Imperfect compartmentalisation may affect neutrophil quantification in crypt-associated inflammation, yet the model performed well for the lamina propria, a key site of infiltration. Overall predictive performance across cohorts indicates that any residual segmentation noise had a limited impact on the biological interpretability of neutrophil density cut-offs, though further refinement will strengthen robustness.

Furthermore, false positives or negatives in AI detection of neutrophils could result in misdiagnosis or overestimation of disease activity and clinical implications. Our AI detected low-level neutrophils in HR-labelled patients, potentially representing true positives missed by pathologists due to sampling bias or subtle staining. Conversely, in challenging instances, the model may misidentify structures with similar morphological features as neutrophils, leading to false positive. Further research is ongoing to refine model architecture. To mitigate this issue and optimise our pipeline, its performance in the PICaSSO cohort was qualitatively evaluated by a second pathologist, that annotated discordant cases.

Notably, that development and validation cohort exhibit different disease activity. The development cohort was purposefully selected for active inflammation to ensure robust neutrophils signals establish a reliable cut-off of treatment response. Conversely, the PICaSSO cohort reflects a more real-world distribution of disease states, including a high proportion of patients in remission with few neutrophils.

Assessing the prognostic value of the model was limited by differences in follow-up data; histology was available in the Mirikizumab cohort but not outcomes, while the validation cohort lacked follow-up histology. Future studies with harmonised data are needed to test its prognostic.

Moreover, in the Mirikizumab cohort there is a large deviation between endoscopic and HR at W12 and W52, as also presented in the original clinical trial,[17] while this difference is negligible in the PICaSSO cohort. However, in the PICaSSO cohort, virtual chromoendoscopy was employed for assessment, better correlating histological findings.[25]

Finally, the relatively small dataset size is a limitation given the complexity of neutrophil analysis. However, it was sufficient for training under the conditions applied, particularly with the use of regularisation techniques including dropout (50%), data augmentation, and early stopping to mitigate overfitting. An external, larger, independent validation cohort would strengthen the generalisability of the findings.

Nonetheless, this novel AI pipeline represents a significant advancement for clinical trials and practice, providing a robust and objective tool for assessing histological activity and therapeutic response in UC. Its application across diverse cohorts, including controlled clinical trials, real-world clinical settings, and active and quiescent patient populations, demonstrates its versatility and adaptability. This innovation can potentially streamline histological evaluations, effort, and cost constraints, and enhance treatment personalisation. While further refinement and extensive validation are ongoing, this approach holds promise for advancing personalised care in UC to improve patient outcomes by bridging the gap between research and clinical applications.

Contributors

MI, IZ and GS have verified the underlying data.

Conceptualisation: MI, VN, EG, SG. Data curation: MI, VV, PM, RD. Formal analysis: VV, PM, RD, DZ, IZ, GS. Investigation: VV, PM, IZ, GS, RD, BK, UC, DZ, BH, RC. Software: VV, PM, RD, VN, EG, BK, UC. Writing—original draft preparation: IZ, GS, VV, PM, RD. Writing—review & editing: MI, DZ, BH, RC, LB, BK, UC, SD, SG, VN, EG. Supervision: MI, LB, SG, SD, EG and VN. All authors read and approved the final version of the manuscript.

Data sharing statement

Data are available from the corresponding author upon reasonable request.

Declaration of interests
MI received research grants and consultant fees from Olympus, Pentax, Eli Lilly, and Janssen. All other authors declare no competing interests.

Acknowledgements
This study was supported by Eli Lilly and Company. We acknowledge Charles Owen, Klaus T. Gottlieb and William J Eastman from Eli Lilly and Company (Indianapolis, IN, USA) for supporting the manuscript preparation.

The PICaSSO group includes Raf Bisschops, Pradeep Bhandari, Gert de Hertogh, Jose G. Ferraz, Martin Goetz, Xianyong Gui, Bu' Hussian Hayee, Ralf Kiesslich, Mark Lazarev, Remo Panaccione, Adolfo Parra-Blanco, Luca Pastorelli, Timo Rath, Vincenzo Villanacci, Elin Synnøve Røyset, Michael Vieth.

Appendix A. Supplementary data
Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103658.

References

1 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.
2 Chateau T, Feakins R, Marchal-Bressenot A, Magro F, Danese S, Peyrin-Biroulet L. Histological remission in ulcerative colitis: under the microscope is the cure. *Am J Gastroenterol*. 2020;115:179–189.
3 Neurath MF, Vieth M. Different levels of healing in inflammatory bowel diseases: mucosal, histological, transmural, barrier and complete healing. *Gut*. 2023;72:2164–2183.
4 Parkes G, Ungaro RC, Danese S, et al. Correlation of mucosal healing endpoints with long-term clinical and patient-reported outcomes in ulcerative colitis. *J Gastroenterol*. 2023;58:990–1002.
5 D'Haens G, Dubinsky M, Kobayashi T, et al. Mirikizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med*. 2023;388:2444–2455.
6 Danese S, Vermeire S, Zhou W, et al. Upadacitinib as induction and maintenance therapy for moderately to severely active ulcerative colitis: results from three phase 3, multicentre, double-blind, randomised trials. *Lancet*. 2022;399:2113–2128.
7 Peyrin-Biroulet L, Bressenot A, Kampman W. Histologic remission: the ultimate therapeutic goal in ulcerative colitis? *Clin Gastroenterol Hepatol*. 2014;12:929–934.e2.
8 Geboes K. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut*. 2000;47:404–409.
9 Nardone OM, Iacucci M, Villanacci V, et al. Real-world use of endoscopic and histological indices in ulcerative colitis: results of a global survey. *United European Gastroenterol J*. 2023;11:514–519.
10 Feakins R, Borralho Nunes P, Driessen A, et al. Definitions of histological abnormalities in inflammatory bowel disease: an ECCO position paper. *J Crohns Colitis*. 2023;18(2):175–191. https://doi.org/10.1093/ecco-jcc/jjad142.
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 Zammarchi I, Santacroce G, Puga-Tejada M, et al. Epithelial neutrophil localization and tight junction Claudin-2 expression are innovative outcome predictors in inflammatory bowel disease. *United European Gastroenterol J*. 2024;12:1155–1166.
13 Iacucci M, Santacroce G, Zammarchi I, et al. Artificial intelligence and endo-histo-omics: new dimensions of precision endoscopy and histology in inflammatory bowel disease. *Lancet Gastroenterol Hepatol*. 2024;9(8):758–772. https://doi.org/10.1016/S2468-1253(24)00053-0.
14 Iacucci M, Parigi TL, Del Amor R, et al. Artificial intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. *Gastroenterology*. 2023;164(7):1180–1188.e2. https://doi.org/10.1053/j.gastro.2023.02.031.
15 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.
16 Ohara J, Maeda Y, Ogata N, et al. Automated neutrophil quantification and histological score estimation in ulcerative colitis. *Clin Gastroenterol Hepatol*. 2024;23(5):846–854.e7. https://doi.org/10.1016/j.cgh.2024.06.040.
17 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.
18 Vadori V, Graïc J-M, Peruffo A, Vadori G, Finos L, Grisan E. CISCA and CytoDArk0: a cell instance segmentation and classification method for histo(patho)logical image analyses and a new, open, Nissl-stained dataset for brain cytoarchitecture studies. *Comput Biol Med*. 2025;197:111018.
19 Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for object detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:936–944.
20 Thiele C, Hirschfeld G. cutpointr: improved estimation and validation of optimal cutpoints in *R*. *J Stat Softw*. 2021;98(11):1–27. https://doi.org/10.18637/jss.v098.i11.
21 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
22 Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. *Gastroenterology*. 2021;160(5):1570–1583. https://doi.org/10.1053/j.gastro.2020.12.031.
23 Mosli MH, Feagan BG, Zou G, et al. Development and validation of a histological index for UC. *Gut*. 2017;66(1):50–58. https://doi.org/10.1136/gutjnl-2015-310393.
24 Marchal-Bressenot A, Salleron J, Boulagnon-Rombi C, et al. Development and validation of the Nancy histological index for UC. *Gut*. 2017;66(1):43–49. https://doi.org/10.1136/gutjnl-2015-310187.
25 Iacucci M, Kiesslich R, Gui X, et al. Beyond white light: optical enhancement in conjunction with magnification colonoscopy for the assessment of mucosal healing in ulcerative colitis. *Endoscopy*. 2017;49(6):553–559. https://doi.org/10.1055/s-0042-124363.