


ORIGINAL RESEARCH

Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density

Peter Bossuyt ,^{1,2} Hiroshi Nakase,³ Séverine Vermeire,¹ Gert de Hertogh,⁴ Tom Eelbode,⁵ Marc Ferrante,¹ Tadashi Hasegawa,⁶ Hilde Willekens,¹ Yousuke Ikemoto,⁷ Takao Makino,⁷ Raf Bisschops¹

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/gutjnl-2019-320056>).

For numbered affiliations see end of article.

Correspondence to

Dr Peter Bossuyt, Department of Gastroenterology, KU Leuven University Hospitals Leuven Gasthuisberg Campus, Leuven B-3000, Belgium; peter.bossuyt@imelda.be

Received 12 October 2019
Revised 5 December 2019
Accepted 29 December 2019
Published Online First
8 January 2020

ABSTRACT

Background The objective evaluation of endoscopic disease activity is key in ulcerative colitis (UC). A composite of endoscopic and histological factors is the goal in UC treatment. We aimed to develop an operator-independent computer-based tool to determine UC activity based on endoscopic images.

Methods First, we built a computer algorithm using data from 29 consecutive patients with UC and 6 healthy controls (construction cohort). The algorithm (red density: RD) was based on the red channel of the red-green-blue pixel values and pattern recognition from endoscopic images. The algorithm was refined in sequential steps to optimise correlation with endoscopic and histological disease activity. In a second phase, the operating properties were tested in patients with UC flares requiring treatment escalation. To validate the algorithm, we tested the correlation between RD score and clinical, endoscopic and histological features in a validation cohort.

Results We constructed the algorithm based on the integration of pixel colour data from the redness colour map along with vascular pattern detection. These data were linked with Robarts histological index (RHI) in a multiple regression analysis. In the construction cohort, RD correlated with RHI ($r=0.74$, $p<0.0001$), Mayo endoscopic subscores ($r=0.76$, $p<0.0001$) and UC Endoscopic Index of Severity scores ($r=0.74$, $p<0.0001$). The RD sensitivity to change had a standardised effect size of 1.16. In the validation set, RD correlated with RHI ($r=0.65$, $p=0.00002$).

Conclusions RD provides an objective computer-based score that accurately assesses disease activity in UC. In a validation study, RD correlated with endoscopic and histological disease activity.

INTRODUCTION

Ulcerative colitis (UC) is a chronic inflammatory disorder involving the colon to different extents.¹ The proposed target for the treatment of UC is complete remission of symptoms combined with endoscopic remission.² It is suggested that a treat-to-target (T2T) algorithm could improve outcomes at the long term in UC.³ But before we can truly

Significance of this study

What is already known on this subject?

- Assessment of disease activity is subjective and leads to interobserver variability.
- Endoscopic scores are useful as treatment target if they are objective and predictive for further disease course.
- Regulatory authorities request a combined endpoint of endoscopic and histological remission for the claim of mucosal healing

What are the new findings?

- Red density (RD) is an operator-independent computer-based tool to determine disease activity in patients with UC.
- RD assesses disease activity based on evaluation of the redness map and vascular pattern recognition.
- RD scores correlated with endoscopic and histological features of UC activity.

How might it impact on clinical practice in the foreseeable future?

- This algorithm might be used for computer analysis of digital endoscopic images from patients with UC and evaluate healing or disease progression in an objective way.
- Larger, prospective studies are ongoing to confirm its accuracy and predictive value.

implement a T2T algorithm, some hurdles have to be taken. First, when using the existing endoscopic scoring systems for UC, it is still to be determined what the exact definition is for endoscopic remission.⁴ In clinical trials, endoscopic remission is defined as Mayo endoscopic subscore (MES) <1 and endoscopic improvement $MES\leq 1$, but in clinical practice, it is suggested that patients with $MES=0$ have longer sustained clinical remission than patient with $MES=1$.⁵ A similar trend is seen when using the Ulcerative Colitis Index of Severity (UCEIS).⁶ Second, all scoring systems are operator dependent with variable interobserver agreement.⁷ In UC, central reading of endoscopic disease activity can



© Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Bossuyt P, Nakase H, Vermeire S, *et al.* *Gut* 2020;**69**:1778–1786.

be of added value in the assessment of a treatment effect and to reduce the placebo effect.⁸ In the reliability and initial validation study of the UCEIS, the interobserver agreement in the determination of UCEIS scores was moderate ($\kappa=0.50$). Additional data demonstrated that even in experienced hands, variable interobserver agreement was seen.⁹ Although training programme can improve the agreement, the subjectivity hampers the use in clinical practice.¹⁰ Third, the significance of a therapeutic target depends on the predictive value for predicting future disease course of the individual patient. Currently, the best predictor for long-term outcomes in patients with UC in clinical remission is histological remission,^{5 11–14} but the scoring is cumbersome.¹⁵ Moreover, a claim of ‘mucosal healing’ requires a combination of endoscopic and histological remission.¹⁶ For this, an endoscopic evaluation tool that is objective, easily accessible and that correlates with endoscopic and histological remission could improve the adoption of a T2T strategy in clinical practice and drug development in UC.

AIM

The aim of the study was to develop an automated real-time operator-independent endoscopic tool that correlates with endoscopic and histological disease activity in UC.

METHODOLOGY

This was a prospective multicentre study performed in two tertiary centres in Belgium and Japan. The study included three phases. In the first phase, the feasibility of the algorithm (‘red density’: RD) was tested, and in several sequential steps, the algorithm was further refined to achieve the optimal correlation with histological and endoscopic disease activity in UC. In the second phase, the operating properties of the RD score developed in phase I were tested in patients with UC that needed treatment escalation for a disease flare. In the third phase, the final algorithm that was developed in phase I was validated in a validation cohort.

1/ System description

We used a high-definition prototype endoscope with a prototype processor from Pentax (Pentax Medical, HOYA Corporation, Tokyo, Japan). White-light (WL) illumination was delivered by a 300W xenon lamp. The processor includes I-scan virtual chromoendoscopy with digital and optical enhancement. The endoscopes had 1290×966 pixels for display. The outer diameter is 9.9 and 11.5 mm with a 105 and 170 cm working length for the gastroscope and colonoscope, respectively. Images were displayed on a 27 inch screen (NDS surgical imaging, San Jose, California, USA). The RD function can be selected from the video processor’s touch panel, the keyboard or the assigned scope button. Once the RD function is activated, RD image and RD score will be displayed on the monitor in real time, along with the WL image (figure 1).

The algorithm used in the study was based on automatic computer-aided assessment of redness on a pixel level. The red channel of the red-green-blue (RGB) pixel values are extracted and used to build an RD map (figure 2A–B)¹⁷: the RD score. For the application of the algorithm, WL images were used before enhancement. The algorithm excludes the interference of stools, light reflection and shadow. The endoscopist needs to wash the mucosa, insufflate air and press the button on the scope to get the RD score, which is comparable to taking a quality endoscopic picture.

2/ Phase I: algorithm development

Consecutive patient with UC and healthy controls (polyp screening) presenting for planned endoscopy at the University Hospitals of Leuven (Leuven, Belgium) and the Sapporo Medical University (Sapporo, Japan) were prospectively recruited in the first phase of the trial. Colon cleansing and sedation was performed according to local guidelines.¹⁸ Endoscopies were performed by two inflammatory bowel disease (IBD) endoscopists with >10 years of experience (PB, HN). The endoscopic procedure followed a standardised protocol (online

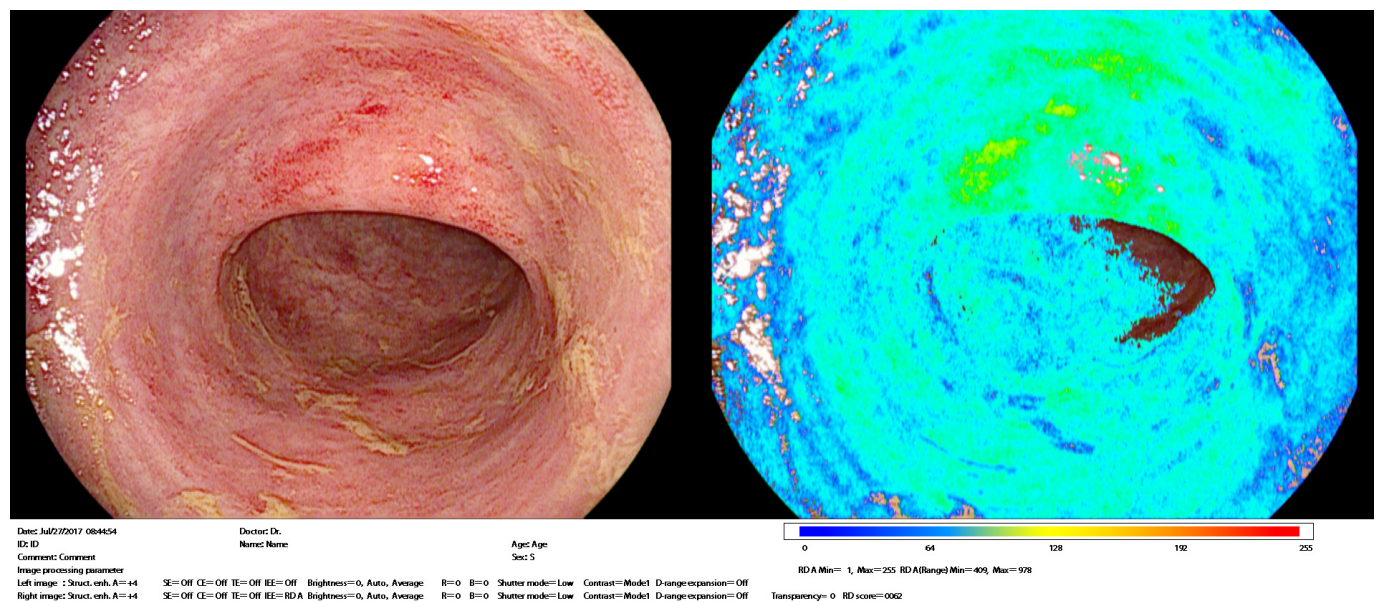


Figure 1 Example of how the red density (RD) system is displayed during endoscopy. RD image and RD score are displayed on the monitor in real time, along with the white light image. The RD image shows the redness of the input image, warmer colour means stronger redness. The RD score is a representative value of colour map and is provided at the bottom of the screen. The colour map and RD score can be recorded simultaneously with still captured image. Left: white light images. Right: RD colour map with RD score in the legend.

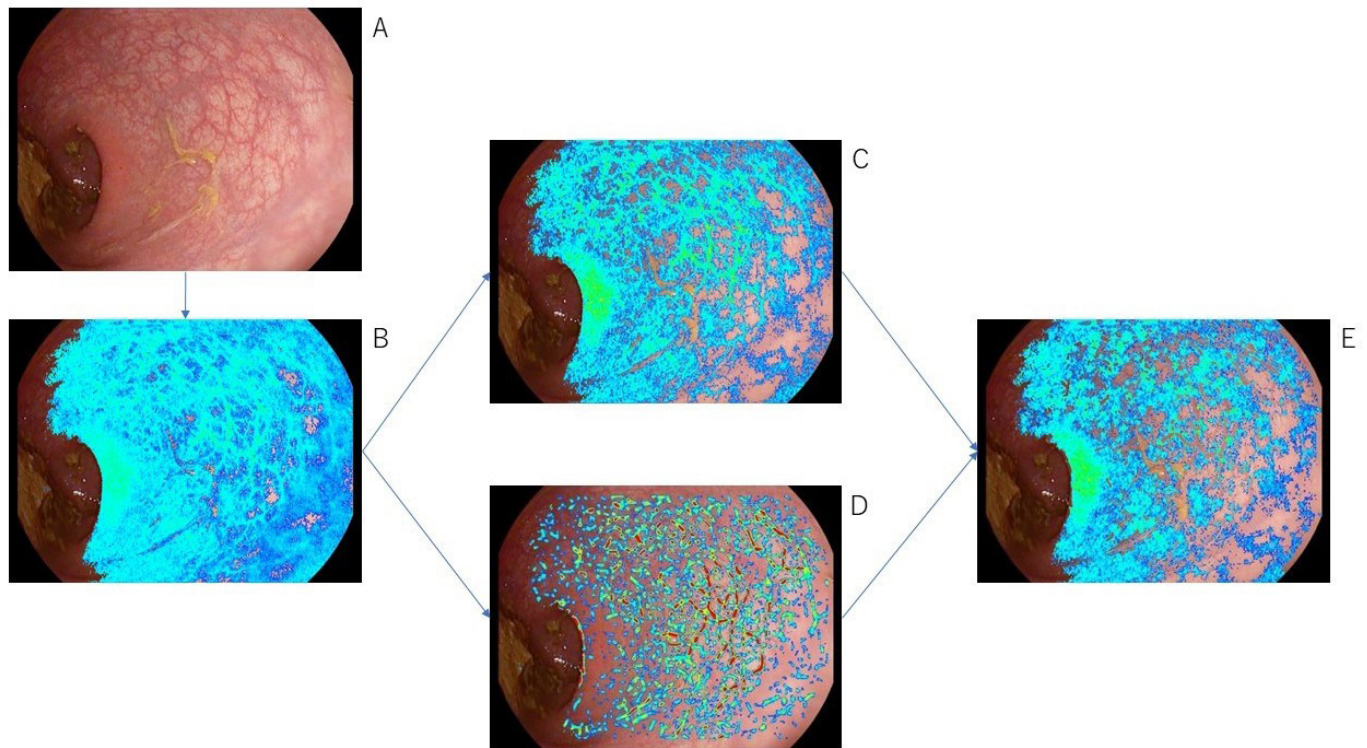


Figure 2 Visual representation of the different modifications in a sample endoscopic image: (A) standard white light high definition endoscopic image; (B) original colour map of the red density image; (C) colour map of the red density image after adapted range setting; (D) image with vascular pattern detection; (E) colour map of the red density after vascular pattern extraction.

supplementary data 1). Biopsies were taken in all evaluated segments both in inflamed (most severe inflammation) and non-inflamed areas (random) of the patients with UC and in healthy controls. All endoscopies were anonymously recorded, and per evaluated segment, still images were obtained. Per segment, all WL images were reassessed at random by two groups of two blinded central readers (SV, HN for the Leuven data and PB, RB for the Sapporo data) for the MES and the UCEIS. The central reader did not assess his own endoscopies he performed before. In case of disagreement, final adjudication was obtained during a consensus meeting with the two central readers. All biopsies were scored according to the Geboes score (GS)¹⁹ and the Robarts histological index (RHI)²⁰ by an experienced UC histopathologist (GDH, TH). Histological remission was defined by different thresholds (GS < 2.0 and GS < 3.1; RHI ≤ 6). Clinical disease activity was recorded based on patient reported outcome (PRO2)²¹ and the total Mayo score.²² Subsequently, the results of the clinical, endoscopic and histological scorings were correlated with the RD score. Based on these initial findings, the algorithm that calculates the RD score was further refined by recalibrating and expanding the RD score, followed by recalculating the RD score in a simulator using still images (Pentax Medical, HOYA Corporation, Tokyo, Japan). By this, it is possible to understand how the RD score is affected by changing the algorithm, and repeated algorithm corrections and adjustments provide optimal results from the input dataset.

3/ Phase II: operating properties testing

Consecutive patients at the University Hospitals of Leuven (Leuven, Belgium) presenting with symptoms of a UC flare for planned endoscopy were included in the trial. Flexible endoscopy was performed before treatment escalation and a second

endoscopic evaluation was performed 8–14 weeks after the treatment escalation to assess endoscopic response and/or remission. At baseline and at week 8–14, PRO2²¹ and total Mayo score were recorded.²² Endoscopic procedure, central reading (SV and MF) and histopathological sampling and scoring (GDH) followed the same protocol as in phase I, except for the fact that videos were used for the central reading of the endoscopic scoring instead of still pictures. This change in format of the central reading compared with the construction phase was done to minimise the inter-rater variability. Only the videos of the most inflamed segment at baseline and the videos of the corresponding segment at week 8–14 were used to test the sensitivity to change.

4/ Phase III: final RD score validation

To validate the final algorithm constructed in phase I, the correlation between RD and clinical, endoscopic and histological scores was retested in the cohort of patients from phase II. In contrast to phase II, all images from all available segments from all patients in phase II were used also of those having only a baseline visit.

5/ Statistical analysis

Statistical analysis was done using R software (The R Foundation, Vienna, Austria). Continuous variables with non-normal distribution are described as medians with IQR. Categorical variables are described as percentages. For time-independent evaluation of continuous variables, we used the Wilcoxon signed-rank test (or Mann-Whitney test if applicable), and for categorical variables, the χ^2 tests (or Fisher's exact test if applicable). Receiver

operating characteristics (ROC) curves were used to determine cut-off values of continuous variables.

Contingency tables took account of the histological remission in relation to RD score, in order to determine the sensitivity, specificity, positive and negative predictive values, and overall diagnostic accuracy of the RD score. The correlations between the RD score and the clinical, endoscopic and histological assessments of disease activity were assessed in Spearman's rank correlation test. The strength of the correlation was described as follows: 0.00–0.19 'very weak'; 0.20–0.39 'weak'; 0.40–0.59 'moderate'; 0.60–0.79 'strong'; 0.80–1.0 'very strong'. The Cohen's kappa was used to calculate the inter-rater agreement between the central readers.

To assess the sensitivity to change of the RD score, we used the standardised effect size.²³ This is based on the difference in the mean RD scores for patients who changed divided by the SD of their baseline scores. Higher values indicate that the evaluative instrument is more responsive to change, and a value >0.8 indicates a large effect.

A formal sample size calculation was not possible for the initial construction cohort, since this was an exploratory exercise to assess correlation between RD and histology. We calculated a sample size with enough sensitivity to detect change. We aimed to show that the RD score significantly decreases after treatment in a population of patients that show clinical benefit (MES2-3 before treatment and MES0-1 after treatment). A sample size of 13 patients, scoring MES2-3 before treatment and MES0-1 after treatment, was needed to show such effect with 80% of power (power increased to 90% when including 16 patients). Calculations are performed for a two-sided paired t-test at 5% significance level, assuming a correlation of 0.05 between both measurements. Using pilot data, we assumed a mean score of 126 and SD 53 before treatment and a mean score after treatment of 68 (SD 43).

RESULTS

1/ Phase I

In total, 29 patients with UC and 6 healthy controls were included in the first phase of the study. Demographic data and disease characteristics of the patient population are summarised in table 1. The development of the RD algorithm started with the data from the initial algorithm based on preliminary testing.¹⁷ The average RD score in the rectum and sigmoid was not significantly different in healthy controls. A higher RD score had more diffuse detection of redness compared with patient with low RD score since redness was only captured in regions with visual vascular pattern, supporting the redness detection. No correlation was seen between the RD score and the haemoglobin level. In the next step, the RD algorithm was tested in 26 patients with UC and 6 healthy controls. There was a correlation when evaluating all segments from all patients (MES: $r=0.38$, $p<0.0001$; UCEIS: $r=0.41$, $p<0.0001$). The RD score was not influenced by C reactive protein level ($r=-0.03$, $p=0.90$) or haemoglobin level ($r=0.26$, $p=0.21$) in patients with UC. In the following step, the range setting for the colour map was revised (figure 2C). By this, the RD score increased overall but more specific in three situations: (1) pictures with high score in the default setting leading to a better discrimination; (2) pictures with a distinct vascular pattern; and (3) pictures with multiple highlights. A vascular pattern recognition algorithm (based on an edge detection technique) was developed to exclude a false-positive effect of vascular pattern (figure 2D). This system automatically detects and scores the presence of vascular structures. An inverse

Table 1 Demographic characteristic of patients in phase I

	Healthy controls (n=6)	UC patients (n=29)
Female gender	5 (83%)	13 (45%)
Age at time of procedure (years): median (IQR)	48.9 (36.3–60.9)	43.6 (38.1–56.7)
Age at diagnosis (years): median (IQR)	NA	32.6 (24.2–43.7)
Disease duration (months): median (IQR)	NA	166 (85–219)
Smoking status	NA	Never 14/27 (52%) Stopped 12/27 (44%) Current 1/27 (4%)
Disease extension (Montreal)	NA	E1: 8/29 (27.6%) E2: 14/29 (48.3%) E3: 7/29 (24.1%)
Current treatment	NA	5-ASA 23/29 (79%) Steroids 3/29 (10%) IMM 3/29 (10%) Anti-TNF 4/29 (14%) Vedolizumab 5/29 (17%)
Previous treatments	NA	5-ASA 29/29 (100%) Steroids 25/29 (86%) IMM 14/29 (48%) Anti-TNF 12/29 (41%) Vedolizumab 4/29 (14%)

5-ASA, 5-aminosalicylic acid; E, extension; IMM, immunomodulator; NA, not applicable; TNF, tumour necrosis factor.

correlation was seen between the vascular pattern score and the vascular subscore of the UCEIS ($r=-0.42$, $p<0.0001$) when evaluating MES 0 and 1 images. In the next step, we applied a new pixel-based analysis by using the 16 bin histograms from the RD colour map (after extraction of the vascular pattern) providing 16 dimensional vectors that were used as 16 explanatory variables in a multiple logistic regression analysis. In that way, the RD score was converted in a probability score ranging from 0 to 200 for the probability of not having MAYO 0. The RD histograms were also used to exclude the interference of ulcers with the vascular pattern score. Subsequently, the converted RD score was combined with the vascular pattern score (figure 2E). This combination provided a correlation with the lower ranges of UCEIS (0–3) ($r=0.54$, $p<0.0001$) and MES (0–1) ($r=0.56$, $p<0.0001$). Since we aimed for an evaluation tool that correlates with histology in UC, we integrated the results of the RHI and the RD histograms of all subjects ($n=70$ images) in a multiple regression analysis; subsequently, the vascular score was integrated in the algorithm. This provided a correlation between the RD score and the final consensus MES ($r=0.71$, $p<0.0001$) and UCEIS ($r=0.69$, $p<0.0001$). The RD score correlates with the RHI ($r=0.60$, $p<0.0001$). To reduce the variation in the RD score, the vascular pattern score setting was further modified and SD of the RD was integrated in the algorithm. The RD histograms were only used to exclude the interference of ulcers. This optimised the correlation between the RD score and RHI ($r=0.61$, $p<0.0001$), MES ($r=0.70$, $p<0.0001$) and UCEIS ($r=0.70$, $p<0.0001$). The calculation of the score based on this algorithm was feasible but necessitated high hardware power. To enable a real-time calculation of the RD score, the algorithm was further modified. For this, the SD of the RD was replaced by the maximum frequency (which negatively correlates with the SD). In this modified setting, the vascular pattern recognition was only used to extract the vascular pattern from the image, not in the formal calculation of the RD score (figure 2E). The final RD

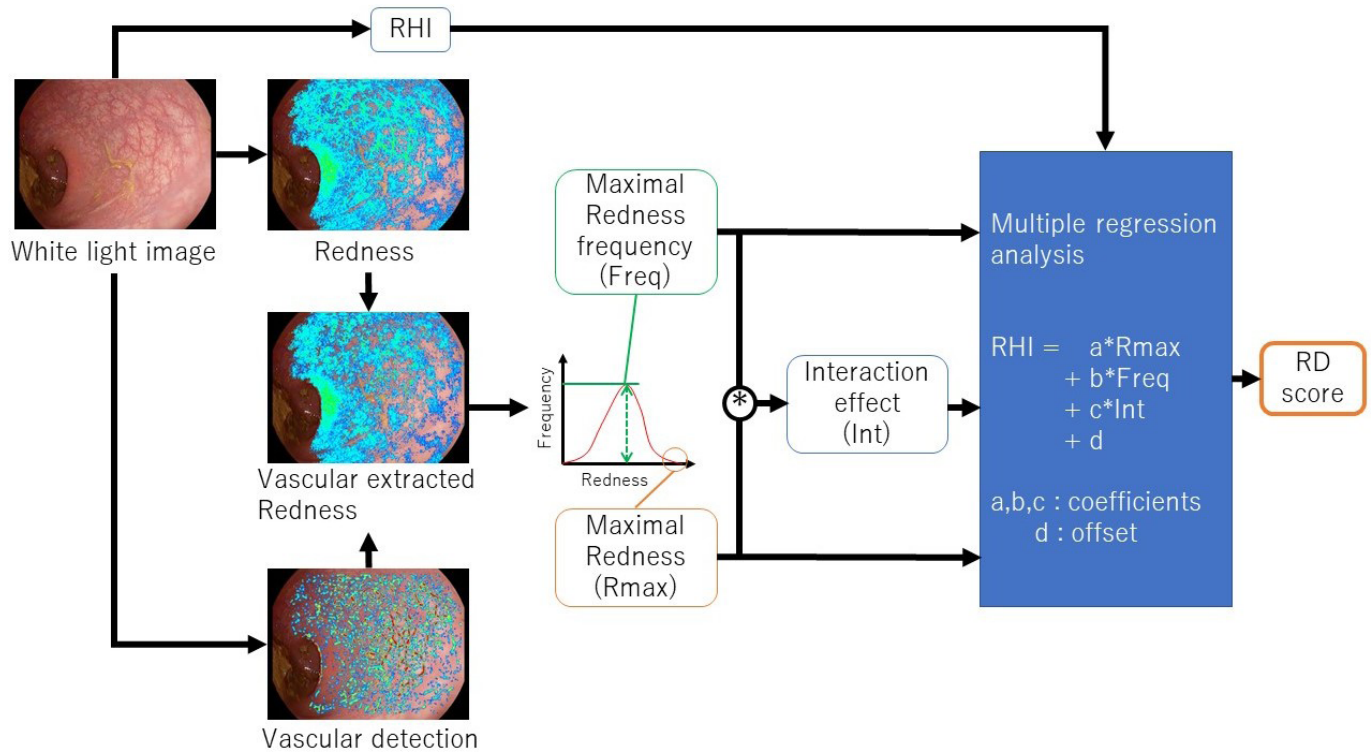


Figure 3 Schematic representation of the construction of the algorithm of the final RD score. RHI, Robarts histological index; Freq, maximal redness frequency; Rmax, maximal redness; Int, interaction effect; RD, red density.

score is based on a multiple regression analysis integrating the maximal redness score, the maximal redness frequency and the RHI of the redness map after extraction of the vascular pattern. The maximal redness is robust to the size of the red area and the interference of ulcers. The final formula of the algorithm is: $RHI = a \cdot Rmax + b \cdot Freq + c \cdot Int + d$. This is schematically represented in figure 3. This results in a final correlation between the RD score and RHI ($r = 0.74$, $p < 0.0001$), MES ($r = 0.76$, $p < 0.0001$) and UCEIS ($r = 0.74$, $p < 0.0001$). The correlation of the total UCEIS was higher than the correlation of the specific subcomponents (vascular: $r = 0.72$; bleeding $r = 0.6$; ulcer $r = 0.61$; all $p < 0.0001$). Based on ROC analysis, with a cut-off RD score of 60, we could discriminate between active histological inflammation ($RHI > 6$) and histological remission ($RHI \leq 6$). An RD score ≤ 60 predicts patient has histological remission ($RHI \leq 6$) with a sensitivity of 96% and specificity of 80%. This results in a positive predictive value for histological remission of 74% and negative predictive value of 97%. The area under the curve (AUC) was 0.95 (online supplementary figure 1). When using other definitions of histological remission based on the GS, a cut-off of 52 and 60 for the RD score was recognised for $GS < 2.0$ and $GS < 3.1$, respectively, see online supplementary figure 2. Inter-rater agreement for the central readers based on still image data was moderate (online supplementary table 1).

2/ Phase II

Sixteen patients were screened and 10 patients were included in the second phase. Demographic data and disease characteristics of the patient population are summarised in table 2. At baseline, the median PRO2 was 3 (IQR 1–4), median total Mayo was 7 (IQR 4–9). All patients had active endoscopic disease at baseline. Median interval between the first and second evaluation was 10 weeks (IQR 8–11). All but one patient received treatment escalation after baseline. The evolution of the clinical,

endoscopic and histological disease activity is shown in figure 4. For central reading, based on the endoscopic videos, adjudication was needed in 50% and 65% of cases for MES and UCEIS, respectively. This resulted in a poor interobserver agreement for both MES ($\kappa = 0.33$) and UCEIS ($\kappa = 0.25$). Nine patients had a change in their endoscopic score compared with baseline. The median delta in UCEIS and MES was 1 (IQR 0–3) ($p = 0.01$) and 1 (IQR 0–1) ($p = 0.003$), respectively. A significant number of patients achieved clinical, endoscopic and histological remission

Table 2 Demographic characteristic of patients in phase II

	UC patients (n=10)
Female gender	6 (60%)
Age at time of first procedure (years): median (IQR)	39.6 (36.3–57.1)
Age at diagnosis (years): median (IQR)	32.2 (23.7–43.8)
Disease duration (months): median (IQR)	76.8 (44.8–144)
Smoking status	Never 5/10 (5%) Stopped 4/10 (40%) Current 1/10 (10%)
Disease extension (Montreal)	E1: 0/10 (0%) E2: 7/10 (70%) E3: 3/10 (30%)
Treatment at baseline	5-ASA 8/10 (80%) Steroids 1/10 (10%) IMM 2/10 (20%) Anti-TNF 1/10 (10%) Vedolizumab 2/10 (20%)
Treatment intensification after baseline	5-ASA 8/10 (80%) Steroids 2/10 (20%) IMM 0/10 (0%) Anti-TNF 1/10 (10%) Vedolizumab 1/10 (10%)

5-ASA, 5-aminosalicylic acid; E, extension; IMM, immunomodulator; NA, not applicable; TNF, tumour necrosis factor.

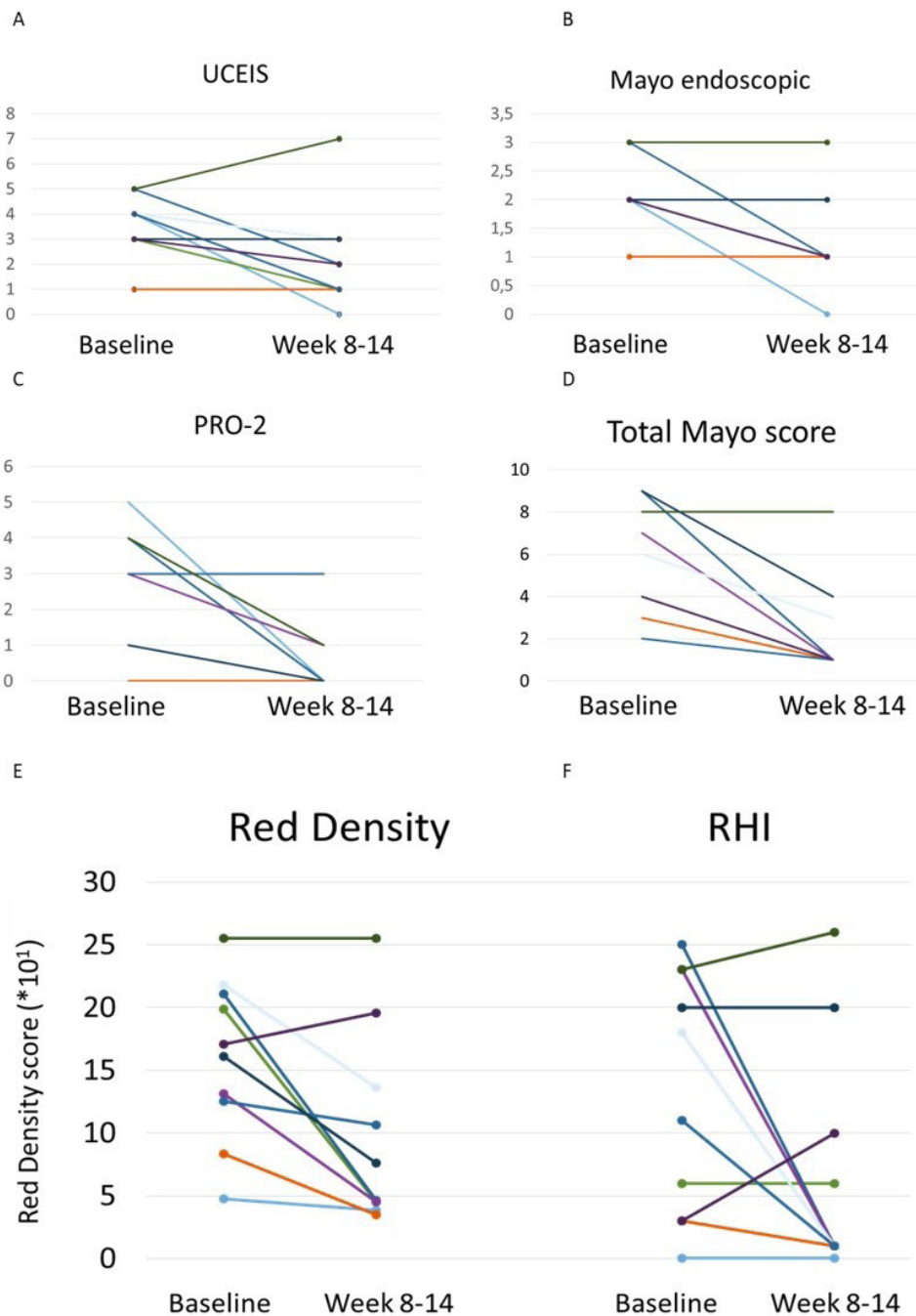


Figure 4 (A–F) Evolution of the endoscopic (A–B) and clinical (C–D) scores, red density score (E) and RHI (F) after treatment escalation. PRO-2, patient-reported outcome 2; RHI, Roberts histological index; UCEIS, Ulcerative Colitis Index of Severity.

after treatment (all $p < 0.03$). Median RD score decreased significantly from baseline (137 to 48; $p = 0.006$) (figure 4E). The standardised effect size for RD was 1.16. There was a correlation between the delta RHI and delta RD ($r = 0.73$; $p = 0.02$). No significant correlation was seen between the delta RD and the delta of MES ($p = 0.4$) or UCEIS ($p = 0.5$).

3/ Phase III

In the final step, the RD score was validated in all available images from the segments of the patients in phase II ($n = 55$). This confirmed the validity of the score and demonstrated a correlation between RD and RHI ($r = 0.65$, $p = 0.00002$) (figure 5). A correlation was seen between UCEIS (based on consensus central

reading ($n = 36$) and the RD score ($r = 0.56$; $p = 0.0004$) and MES ($r = 0.61$; $p = 0.00009$), but there was a wide range of RD values for MES=1 and UCEIS=1, which underlines the subjectivity of the interpretation of MES and UCEIS by the human eye (online supplementary figure 3).

DISCUSSION

The RD score is an important new concept in the evaluation of disease activity in UC. It is the first endoscopic scoring system that provides operator-independent full digital scoring of disease activity in UC. Due to the integration of pattern recognition and automated redness assessment from the endoscopic images, it excludes the subjectivity of the operator. RD correlates

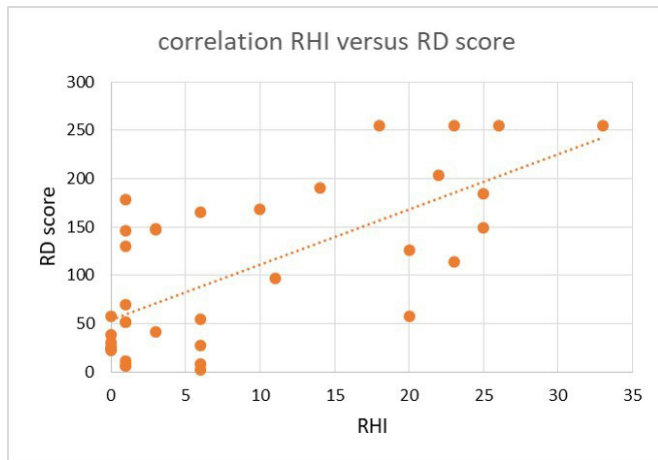


Figure 5 Correlation between RHI and RD score for the patients in phase II ($r=0.65$, $p=0.00002$). RD, red density; RHI, Robarts histological index.

with histological scoring systems (RHI) and to a lower extent with existing subjective endoscopic scores. For this, RD evaluates endoscopic images on a deeper level. Histology is a better predictor of further disease course in UC and endoscopic scores are hampered by inter-rater and intra-rater variability. In this way, there is a potential for RD in the prediction of sustained remission and a target in a T2T setting.³ RD demonstrates an excellent sensitivity to change. RD can be used as an objective monitoring tool for the evaluation of treatment effect in UC.

Recently, several endoscopic innovations for the scoring of disease activity in UC have been proposed. The Picasso score is based on high-definition endoscopic images and electronic virtual chromoendoscopy.²⁴ Due to its superb image quality, vascular dilation, mucosal architecture and intramucosal bleeding can be appraised. Although specific training programme for the Picasso score results in low inter-rater variability in expert hands, data on the utility for real-time use in daily clinical practice are lacking.²⁵ Confocal laser endomicroscopy (CLE) deals with similar problems as other endoscopic scores in IBD. This technique performs well in expert hands, but has a certain learning curve. Moreover, CLE requires intravenous fluorescein administration and specific endoscopic equipment.^{26,27} RD, in contrast, can be used without specific training. The use of RD does not require special preparation and the scoring results are provided in real time.

Redness is difficult to score objectively by the human eye. Uchiyama *et al* recently used a linked colour imaging (LCI) index based on postprocessing of small region of interest in the evaluation of endoscopic healing in UC.²⁸ They demonstrated a correlation between histopathology and the LCI index. RD in contrast is constructed based on the complete endoscopic image and a multifaceted approach with redness data per pixel and an integrated vascular pattern recognition.

With the introduction of machine learning and pattern recognition in endoscopy, a new era has begun. The use of neural networks for computer-aided diagnosis is promising for the evaluation and classification of colonic polyps.^{29,30} Two groups reported excellent performance of neural network to assess endoscopic disease activity based on the MES.^{31,32} Although these results seem promising, nevertheless this deep learning system is constructed based on the subjective evaluation of endoscopists who provide the scoring of the images that trained the convolutional neural network. Deep learning is used in almost all fields of computer vision and with good

reason: its performance surpasses many of the more traditional computer vision techniques (sometimes even the human performance)³³ and it is extremely fast. There are, however, still several limitations to this technique which can justify the use of other machine learning algorithms like RD. First of all, training a deep neural network requires generally a very large annotated dataset before it can generalise well to the given data. The failure modes of a deep neural network are unpredictable and its results are uninterpretable as to how it got to its conclusion. This might not be an issue for many computer vision tasks, but can be different for automated clinical diagnosis using clinical images. So, if there is a task with limited data and we care about having a predictable and explainable outcome from the algorithm, one might prefer to go with some of the more 'traditional methods' like feature extracting clinical images. In this context, the RD score overcomes this limitation by building the score based on objective imaging data that were correlated with histological scoring.

Our study has some limitations. First, in the initial phase of the study, predominantly patients with low disease activity were included. However, for endoscopists the challenge lays mainly in discrimination MES 1 from MES 0 and this part of the spectrum is the target in a T2T setting in UC. Second, the total number of patients that were included in the study was modest ($n=45$). But since we assessed at least 10 images per patients and tested the algorithm in a simulator, an infinite number of testing rounds could be done. In contrast to other systems for computer-aided diagnosis like convolutional neural networks that require thousands of images, our approach needs a significantly lower amount of data due to the possibility of sequential modulation of the algorithm during the development. Furthermore, the current number of endoscopies used for the construction of RD is in line with what is conventional for scores based on human evaluation like UCEIS ($n=60$).³⁴ RD is an automated algorithm, excluding inter-rater variability and providing always the same score for the same image; for this, multiple reading and re-reading during construction is not required. Third, the algorithm works for still images and currently does not work yet for moving images. But all current scoring systems and predictive models in UC are based on the dominant score of the most affected segment. For this reason, RD cannot also be used in patients with Crohn's disease due to the irregular distribution of the disease and the use of non-dominant cumulative scoring systems for endoscopic disease activity. At this stage, we did not correlate or integrate the RD score with other disease biomarkers like faecal calprotectin, but this will be done in a follow-up prospective validation study. We wanted first to correlate the RD score with histology since its value is a predictor of disease course in UC. Last, in this stage the system is only available for Pentax prototype endoscopes. By adapting the algorithm to other systems, it might have the potential to be used in different endoscopy platforms, although this needs to be developed and confirmed first.

The RD score needs further clinical validation. A multicentre study in patients with UC in clinical remission is planned to assess the predictive value of the RD score for sustained clinical remission. If this multicentre study confirms a cut-off value of the RD score that predicts favourable long-term outcomes in UC, then the RD score can be used as the first objective operator-independent endoscopic target in a T2T strategy in UC.²

In conclusion, we developed the first objective operator-independent endoscopic scoring system for disease activity in UC with excellent operating properties and correlation with both endoscopic and histological disease activity.

Author affiliations

¹Department of Gastroenterology and Hepatology, University Hospitals Leuven, KU Leuven, Leuven, Belgium

²Department of Gastroenterology, Imelda GI Clinical Research Centre, Imelda General Hospital, Bonheiden, Belgium

³Department of Gastroenterology, Sapporo Medical University, Sapporo, Japan

⁴Department of Pathology, University Hospitals Leuven, KU Leuven, Leuven, Belgium

⁵Medical Imaging Research Center, University Hospitals Leuven, KU Leuven, Leuven, Belgium

⁶Department of Surgical Pathology, Sapporo Medical University, Sapporo, Japan

⁷Product Development Department, Pentax Medical, Tokyo, Japan

Twitter Peter Bossuyt @ibd_kliniek

Contributors HN developed the concept of red density (RD) based on the Magic score system. HN, PB and RB contributed equally to the concept of the study and the further development of the RD system in this study. PB, HN, GdH, MF, TH, SV and RB: acquisition of data; PB, YI, TM, RD: analysis and interpretation of data; PB: drafting of the manuscript; all: critical revision of the manuscript for important intellectual content; PB, TM: statistical analysis; RB: obtained funding; HW, YI and TM administrative, technical or material support; RB: study supervision.

Funding This project was supported by an unrestricted grant from Pentax. The authors codeveloped the system where the clinical input came from the authors' side and Pentax provided the technical input to fine-tune the algorithm. As clinical researchers, the authors are not involved with the potential commercialisation of this technique.

Competing interests PB has received financial support for research from AbbVie, Mundipharma, Pfizer, Janssen and Mylan; lecture fees from AbbVie, Takeda, Pfizer and Janssen; advisory board fees from AbbVie, Takeda, Hospira, Janssen, MSD, Mundipharma, Roche, Pfizer, Sandoz and Pentax. HN has received advisory board fees from Kyorin Pharmaceutical, Mochida Pharmaceutical, Janssen Pharmaceutical K.K., Pfizer; lecture fee from: Janssen Pharmaceutical K.K., Pfizer, Takeda Pharmaceutical, Mitsubishi Tanabe Pharma (MTP), Abbvie GK(AGK), Eisai Corporation (EC), Kyorin Pharmaceutical, Zeria Pharmaceutical, Mochida Pharmaceutical, Janssen Pharmaceutical K.K., Nippon Kayaku; commissioned/joint research grant: Hoya Group Pentax Medical, Boehringer Ingelheim GmbH. SV has received financial support for research: MSD, AbbVie, Takeda, Pfizer, J&J; Lecture fee(s): MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche; Consultancy: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche, Celgene, Mundipharma, Celltrion, SecondGenome, Promethues, Shire, Prodigest, Gilead, Galapagos, MRM Health. GdH's institution KULeuven received fees for his activities as central pathology reviewer in clinical trials of Centocor and Takeda. TE reports no conflict of interest. MF has received research grants: Janssen, Pfizer, Takeda; consultancy fees: Abbvie, Boehringer-Ingelheim, Celltrion, Ferring, Janssen, Lilly, Mitsubishi Tanabe, MSD, Pfizer, Takeda; speakers fees: Abbvie, Amgen, Biogen, Boehringer-Ingelheim, Chiesi, Falk, Ferring, Janssen, Lampro, Mitsubishi Tanabe, MSD, Pfizer, Takeda, Tramedico, Tillotts and Zeria. HW reports no conflict of interest. YI is an employee of Pentax Medical. TM is an employee of Pentax Medical. RB has received research grant consultancy and speaker's fees: Pentax Medical and Fujifilm.

Patient consent for publication Not required.

Ethics approval The study was conducted in accordance with the ethical principles stated in the Declaration of Helsinki 2008 and in compliance with the principles of Good Clinical Practice (GCP), according to the International Conference on Harmonisation Harmonised Tripartite Guideline. Prior to initiation of the study at any site, the study, including the protocol, informed consent, and other study documents, was approved by an appropriate Independent Ethics Committee of the participating centres (s59405 EC approval number Leuven, 282-185 EC approval number Sapporo).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data may be obtained from a third party and are not publicly available. All data relevant to the study are included in the article or uploaded as supplementary information. All data are deidentified clinical data from participant. Mathematic details on the algorithm are not included but are available on reasonable request. Technical details on the endoscope system need to be obtained by a third party (Pentax Medical) and are not publicly available.

ORCID iD

Peter Bossuyt <http://orcid.org/0000-0003-4027-7365>

REFERENCES

1 Ungaro R, Mehandru S, Allen PB, *et al.* Ulcerative colitis. *The Lancet* 2017;389:1756–70.

- 2 Peyrin-Biroulet L, Sandborn W, Sands BE, *et al.* Selecting therapeutic targets in inflammatory bowel disease (STRIDE): determining therapeutic goals for Treat-to-Target. *Am J Gastroenterol* 2015;110:1324–38.
- 3 Bossuyt P, Vermeire S. Treat to target in inflammatory bowel disease. *Gastrointest Endosc Clin N Am* 2019;29:421–36.
- 4 Vuitton L, Peyrin-Biroulet L, Colombel JF, *et al.* Defining endoscopic response and remission in ulcerative colitis clinical trials: an international consensus. *Aliment Pharmacol Ther* 2017;45:801–13.
- 5 Ponte A, Pinho R, Fernandes S, *et al.* Impact of histological and endoscopic remissions on clinical recurrence and recurrence-free time in ulcerative colitis. *Inflamm Bowel Dis* 2017;23:2238–44.
- 6 Ikeya K, Hanai H, Sugimoto K, *et al.* The ulcerative colitis endoscopic index of severity more accurately reflects clinical outcomes and long-term prognosis than the Mayo endoscopic score. *ECCOJC* 2016;10:286–95.
- 7 Daperno M, Comberlato M, Bossa F, *et al.* Inter-Observer agreement in endoscopic scoring systems: preliminary report of an ongoing study from the Italian group for inflammatory bowel disease (IG-IBD). *Dig Liver Dis* 2014;46:969–73.
- 8 Feagan BG, Sandborn WJ, D'Haens G, *et al.* The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology* 2013;145:149–57.
- 9 Travis SPL, Schnell D, Krzeski P, *et al.* Reliability and initial validation of the ulcerative colitis endoscopic index of severity. *Gastroenterology* 2013;145:987–95.
- 10 Daperno M, Comberlato M, Bossa F, *et al.* Training programs on endoscopic scoring systems for inflammatory bowel disease lead to a significant increase in interobserver agreement among community Gastroenterologists. *J Crohns Colitis* 2017;11:556–61.
- 11 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;65:408–14.
- 12 Zenlea T, Yee EU, Rosenberg L, *et al.* Histology grade is independently associated with relapse risk in patients with ulcerative colitis in clinical remission: a prospective study. *Am J Gastroenterol* 2016;111:685–90.
- 13 Lobatón T, Bessissow T, Ruiz-Cerulla A, *et al.* Prognostic value of histological activity in patients with ulcerative colitis in deep remission: a prospective multicenter study. *United European Gastroenterol J* 2018;6:765–72.
- 14 Bessissow T, Lemmens B, Ferrante M, *et al.* Prognostic value of serologic and histologic markers on clinical relapse in ulcerative colitis patients with mucosal healing. *Am J Gastroenterol* 2012;107:1684–92.
- 15 Römkens TEH, Kranenburg P, Tilburg Avan, *et al.* Assessment of histological remission in ulcerative colitis: discrepancies between daily practice and expert opinion. *J Crohns Colitis* 2018;12:425–31.
- 16 United States food and drugs administration. Ulcerative colitis: clinical trial endpoints guidance for industry, 2016. Available: <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM515143.pdf>
- 17 Higuchi H, Yoshino T, Matsuura M, *et al.* Su1115 a novel endoscopic imaging system for evaluation of colonic mucosal inflammation in clinically quiescent ulcerative colitis. *Gastroenterology* 2013;144:S-403.
- 18 Hassan C, Bretthauer M, Kaminski M, *et al.* Bowel preparation for colonoscopy: European Society of gastrointestinal endoscopy (ESGE) guideline. *Endoscopy* 2013;45:142–55.
- 19 Geboes K, Riddell R, Ost A. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut* 2000;47:404–9.
- 20 Mosli MH, Feagan BG, Zou G, *et al.* Development and validation of a histological index for UC. *Gut* 2017;66:50–8.
- 21 Jairath V, Khanna R, Zou GY, *et al.* Development of interim patient-reported outcome measures for the assessment of ulcerative colitis disease activity in clinical trials. *Aliment Pharmacol Ther* 2015;42:1200–10.
- 22 Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. *N Engl J Med* 1987;317:1625–9.
- 23 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures statistics and strategies for evaluation. *Control Clin Trials* 1991;12:S142–58.
- 24 Iacucci M, Daperno M, Lazarev M, *et al.* Development and reliability of the new endoscopic virtual chromoendoscopy score: the PICA^{SSO} (Paddington International Virtual ChromoendoScopy ScOre) in ulcerative colitis. *Gastrointest Endosc* 2017;86:1118–27.
- 25 Trivedi PJ, Kiesslich R, Hodson J, *et al.* The Paddington international virtual Chromoendoscopy score in ulcerative colitis exhibits very good inter-rater agreement after computerized module training: a multicenter study across academic and community practice (with video). *Gastrointest Endosc* 2018;88:95–106.
- 26 Chang J, Ip M, Yang M, *et al.* The learning curve, interobserver, and intraobserver agreement of endoscopic confocal laser endomicroscopy in the assessment of mucosal barrier defects. *Gastrointest Endosc* 2016;83:785–91.
- 27 Macé V, Ahluwalia A, Coron E, *et al.* Confocal laser endomicroscopy: a new gold standard for the assessment of mucosal healing in ulcerative colitis. *J Gastroenterol Hepatol* 2015;30:85–92.
- 28 Uchiyama K, Takagi T, Kashiwagi S, *et al.* Assessment of endoscopic mucosal healing of ulcerative colitis using linked colour imaging, a novel endoscopic enhancement system. *J Crohns Colitis* 2017;11:963–9.

- 29 Chen P-J, Lin M-C, Lai M-J, *et al.* Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;154:568–75.
- 30 Byrne MF, Chapados N, Soudan F, *et al.* Real-Time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut* 2019;68:94–100.
- 31 Ozawa T, Ishihara S, Fujishiro M, *et al.* Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest Endosc* 2019;89:416–21.
- 32 Stidham RW, Liu W, Bishu S, *et al.* Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open* 2019;2:e193963.
- 33 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
- 34 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;61:535–42.