Title: Automatic, Computer-aided Determination of Endoscopic and Histologic Inflammation in Patients With Mild To Moderate Ulcerative Colitis Based on Red Density

Short Title: Red density: automated endoscopic scoring.

Authors: Peter Bossuyt ^{1,2}, Hiroshi Nakase ³, Séverine Vermeire ¹, Gert De Hertogh ⁴, Tom Eelbode ⁵, Marc Ferrante¹, Tadashi Hasegawa ⁶, Hilde Willekens ¹, Yousuke Ikemoto ⁷, Takao Makino ⁷, Raf Bisschops ¹.

Author Affiliations:

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

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

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

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

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

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

⁷ Pentax Medical, Product development department, Tokyo, Japan

Corresponding author:

Peter Bossuyt, M.D.

Department of Gastroenterology and Hepatology

University hospitals Leuven, KU Leuven

Herestraat 49

B-3000 Leuven, BELGIUM

Tel : 0032 15 50 51 65

Email: peter.bossuyt@imelda.be

Author Contributions: HN developed the concept of 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.

Disclosures:

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 CO.,Ltd., Mochida Pharmaceutical CO.,Ltd., Janssen Pharmaceutical K.K., Pfizer Lnc; lecture fee from: Janssen Pharmaceutical K.K., Pfizer Lnc, Takeda Pharmaceutical CO.,Ltd., Mitsubishi Tanabe Pharma (MTP) CO.,Ltd., Abbvie GK(AGK), Eisai Corporation (EC), Kyorin Pharmaceutical CO.,Ltd., Zeria Pharmaceutical CO.,Ltd., Mochida Pharmaceutical CO.,Ltd., Janssen Pharmaceutical K.K, Nippon Kayaku Co.,Ltd.; 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, Prometheus, 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, Zeria

TH reports no conflict of interest

HW reports no conflict of interest

YI employee of Pentax Medical

TM employee of Pentax Medical

RB has received research grant consultancy and speaker's fees: Pentax Medical and Fujifilm

Abbreviations:

5-ASA: 5-aminosalicylic acid

AUC: area under the curve

CEI: commission internationale de l'éclairage

CLE: confocal laser endomicroscopy

E: extension

FDA: food and drug administration

Freq: Maximal Redness frequency

GS: Geboes score

IBD: inflammatory bowel disease

IMM: immunomodulatory

Int: Interaction effect

IQR: interquartile range

LCI: linked color imaging

MES: Mayo endoscopic subscore

NA: not applicable

PRO: patient reported outcomes

RD: red density

RGB: red green blue

RHI: Robarts histological index

Rmax: Maximal Redness

ROC: receiver operating characteristics

SD: standard deviation

T2T: treat-to-target

TNF: tumor necrosis factor

UC: ulcerative colitis

UCEIS: ulcerative colitis index of severity

WL: white-light

Grant support: this project was supported by an unrestricted grant from Pentax. The authors co-developed 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 commercialization of this technique.

Writing Assistance: none

Word count main text: 4245 words

Word count abstract: 250 words

Figures count: 5

Tables count: 2

Supplementary figures: 3

Supplementary tables: 1

Abstract:

Background: The objective evaluation of endoscopic disease activity is key in ulcerative colitis (UC). A composite of endoscopic and histologic 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 optimize 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 histologic features in a validation cohort.

Results: We constructed the algorithm based on the integration of pixel color data from the redness color map along with vascular pattern detection. These data were linked with Robarts histologic 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 standardized effect size of 1.16. In the validation set, RD correlated with RHI ($r=0.65$, $p=0.00002$).

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

KEY WORDS: artificial intelligence, evaluation, response to treatment, remission

Summary Box:

What is already known about 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 is an operator-independent computer-based tool to determine disease activity in patients with ulcerative colitis (UC).
- Red Density assesses disease activity based on evaluation of the redness map and vascular pattern recognition.
- Red Density scores correlated with endoscopic and histologic features of ulcerative colitis 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.

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 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 inter-observer agreement. [7] In UC central reading of endoscopic disease activity can 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 inter-observer agreement in the determination of UCEIS scores was moderate ($\kappa = 0.50$). Additional data demonstrated that even in

experienced hands variable inter-observer agreement was seen. [9] Although training programs 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 UC patients in clinical remission is histological remission, [5, 11, 12, 13, 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 multicenter study performed in two tertiary center 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 a second phase, the operating properties of the red density score developed in phase 1 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 1 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 chromo-endoscopy with digital and optical

enhancement. The endoscopes had 1290x966 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, CA, 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 white light 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 a red density map (figure 2a-b) [17]: the red density 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 1: 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 >10y of experience (PB, HN). The endoscopic procedure followed a standardized protocol (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 re-assessed ad 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 assessed 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 $GS < 3.1$; $RHI \leq 6$). 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 to 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 data set.

3/ phase 2: 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 to 14 weeks after the treatment escalation to assess endoscopic response and/or remission. At baseline and at week 8-14, PRO2 [21] and total Mayo score was recorded [22]. Endoscopic procedure, central reading (SV and MF) and histopathological sampling and scoring (GDH) followed the same protocol as in phase 1, 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 to the construction phase was done to minimize 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 3: final RD score validation

To validate the final algorithm constructed in phase 1, the correlation between RD and clinical, endoscopic and histological scores was retested in the cohort of patients from phase 2. In contrast to

phase 2, all images from all available segments from all patients in phase 2 were used also of those having only a baseline visit.

5/ ethical aspects

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 Harmonization Harmonized 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 centers (s59405 EC approval number Leuven, 282-185 EC approval number Sapporo).

6/ 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 inter-quartile range (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 chi-square 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 followed: 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 standardized effect size [23]. This is based on the difference in the mean RD scores for patients who changed divided by the standard deviation (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 red density 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

Phase 1

In total 29 patients with ulcerative colitis and 6 healthy controls were included in the first phase of the study. Demographic data and disease characteristics of the patient population are summarized 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 to 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 hemoglobin 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 hemoglobin level ($r = 0.26$, $p = 0.21$) in patients with UC. In a following step the range setting for the color map were revised (figure 2c). By this, the RD score increased overall but more specific in three situations: (i) pictures with high score in the default setting leading to a better discrimination. (ii) pictures with a distinct vascular pattern and (iii) 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 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 a next step we applied a new pixel-based analysis by using the 16 bin histograms from the RD color 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 range 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 were further modified and standard deviation of the RD was integrated in the algorithm. The RD histograms were only used to exclude the interference of ulcers. This optimized the correlation between the RD score and RHI ($r = 0.61$, $p < 0.0001$), MES ($r = 0.70$, $p < 0.0001$), 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 standard deviation of the RD was replaced by the maximum frequency (which negatively correlates with the standard deviation). 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 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: $RHI = a \cdot R_{max} + b \cdot Freq + c \cdot Int + d$. 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 ($r=0.76$, $p<0.0001$), 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 a cut off of RD score of 60 we could discriminate between active histological inflammation ($RHI>6$) and histological remission ($RHI\leq 6$). A 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 (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 recognized respectively for $GS<2.0$ and $GS<3.1$, see supplementary figure 2. Inter-rater agreement for the central readers based on still image data was moderate (supplementary table 1).

Phase 2

Sixteen patients were screened and 10 patients were included in the second phase. Demographic data and disease characteristics of the patient population are summarized 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 inter-observer agreement for both MES ($\kappa = 0.33$) and UCEIS ($\kappa = 0.25$). Nine patients had a change in their endoscopic score compared to 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 after treatment (all $p < 0.03$). Median RD score decreased significantly from baseline (137 to 48; $p=0.006$) (figure 4e). The standardized 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 ($p=0.5$).

Phase 3

In a final step the RD score was validated in all available images from the segments of the patients in phase 2 ($n=55$). 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 (supplementary figure 3).

Discussion

The red density 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 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 for further disease course in UC and endoscopic scores are hampered by inter- and intra-rater variability. In this way there is a potential for RD in the prediction of sustained remission and a target in a treat-to-target 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 chromo-endoscopy. [24] Due to its superb image quality, vascular dilation, mucosal architecture and intra-mucosal bleeding can be appraised. Although specific training programs for the Picasso score result 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 requires 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 color imaging (LCI) index based on post-processing 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: it's 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 generalize 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 of 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 treat-to-target setting in UC. Second, the total number of patients 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 the fact 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 not 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 also can't 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 fecal 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 as 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 multicenter study in patients with UC in clinical remission is planned to assess the predictive value of the RD score for sustained clinical remission. If an objective cut-off value of the RD score can be confirmed that predicts favorable long term outcomes in UC then the RD score can be used as the first objective operator independent endoscopic target in a treat-to-target 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.

References

- 1 Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel JF. Ulcerative colitis. *Lancet* 2017;**389**:1756-70.
- 2 Peyrin-Biroulet L, Sandborn W, Sands BE, Reinisch W, Bemelman W, Bryant RV, *et al*. Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE): Determining Therapeutic Goals for Treat-to-Target. *Am J Gastroenterol* 2015.
- 3 Bossuyt P, Vermeire S. Treat to Target in Inflammatory Bowel Disease. *Curr Treat Options Gastroenterol* 2016.

- 4 Vuitton L, Peyrin-Biroulet L, Colombel JF, Pariente B, Pineton de Chambrun G, Walsh AJ, *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, Rodrigues A, Alberto L, Silva JC, *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, Osawa S, Kawasaki S, Iida T, *et al.* The Ulcerative Colitis Endoscopic Index of Severity More Accurately Reflects Clinical Outcomes and Long-term Prognosis than the Mayo Endoscopic Score. *J Crohns Colitis* 2016;**10**:286-95.
- 7 Daperno M, Comberlato M, Bossa F, Biancone L, Bonanomi AG, Cassinotti A, *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, Pola S, McDonald JW, Rutgeerts P, *et al.* The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology* 2013;**145**:149-57.e2.
- 9 Travis SP, Schnell D, Krzeski P, Abreu MT, Altman DG, Colombel JF, *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, Armuzzi A, Biancone L, Bonanomi AG, *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, Walsh AJ, Thomas S, von Herbay A, *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, Boyle M, Nanda KS, Wolf JL, *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 Lobaton T, Bessissow T, Ruiz-Cerulla A, De Hertogh G, Bisschops R, Guardiola J, *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, Bisschops R, Van Steen K, Geboes K, *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 Romkens TEH, Kranenburg P, Tilburg AV, Bronkhorst C, Nagtegaal ID, Drenth JPH, *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. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM515143.pdf>, 2016.
- 17 Higuchi H, Yoshino T, Matsuura M, Nakase H, Chiba T. 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 MF, Polkowski M, Rembacken B, Saunders B, *et al.* Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) guideline. *Endoscopy* 2013;**45**:142-50.
- 19 Geboes K, Riddell R, Ost A, Jensfelt B, Persson T, Lofberg R. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut* 2000;**47**:404-9.
- 20 Mosli MH, Feagan BG, Zou G, Sandborn WJ, D'Haens G, Khanna R, *et al.* Development and validation of a histological index for UC. *Gut* 2017;**66**:50-8.

- 21 Jairath V, Khanna R, Zou GY, Stitt L, Mosli M, Vandervoort MK, *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. A randomized study. *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**:142s-58s.
- 24 Iacucci M, Daperno M, Lazarev M, Arsenescu R, Tontini GE, Akinola O, *et al.* Development and reliability of the new endoscopic virtual chromoendoscopy score: the PICaSSO (Paddington International Virtual ChromoendoScopy ScOre) in ulcerative colitis. *Gastrointest Endosc* 2017;**86**:1118-27.e5.
- 25 Trivedi PJ, Kiesslich R, Hodson J, Bhala N, Boulton RA, Cooney R, *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.e2.
- 26 Chang J, Ip M, Yang M, Wong B, Power T, Lin L, *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.e1.
- 27 Mace V, Ahluwalia A, Coron E, Le Rhun M, Boureille A, Bossard C, *et al.* Confocal laser endomicroscopy: a new gold standard for the assessment of mucosal healing in ulcerative colitis. *J Gastroenterol Hepatol* 2015;**30 Suppl 1**:85-92.
- 28 Uchiyama K, Takagi T, Kashiwagi S, Toyokawa Y, Tanaka M, Hotta Y, *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 PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS. Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. *Gastroenterology* 2018;**154**:568-75.
- 30 Byrne MF, Chapados N, Soudan F, Oertel C, Linares Perez M, Kelly R, *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, Saito H, Kumagai Y, Shichijo S, *et al.* Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest Endosc* 2018.
- 32 Stidham RW, Liu W, Bishu S, Rice MD, Higgins PDR, Zhu J, *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, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**:115-8.
- 34 Travis SP, Schnell D, Krzeski P, Abreu MT, Altman DG, Colombel JF, *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.

Figures

Figure 1: example of how the red density system is displayed during endoscopy. Red density image and red density score are displayed on the monitor in real time, along with the white light image. The red density image shows the redness of the input image, warmer color means stronger redness. The

RD score is a representative value of color map and is provide at the bottom of the screen. The color map and RD score can be recorded simultaneously with still captured image. Left: white light images. Right: red density color map with red density 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 color map of the red density image; (C) color map of the red density image after adapted range setting; (D) image with vascular pattern detection; (E) color map of the red density after vascular pattern extraction.

Figure 3: Schematic representation of the construction of the algorithm of the final red density score. (RHI: Robarts histological index, Freq: Maximal Redness frequency, Rmax: Maximal Redness, Int: Interaction effect, RD: red density)

Figure 4 a-f: evolution of the endoscopic (a-b) and clinical (c-d) scores, red density score (e) and Robarts histological index (f) after treatment escalation. (UCEIS: ulcerative colitis index of severity; PRO-2: patient reported outcome 2, RHI: Robarts histological index)

Figure 5: Correlation between Robarts histological index and red density score for the patients in phase 2 ($r=0.65$, $p=0.00002$). (RHI: Robarts histological index; RD: red density)

Tables:

Table 1: demographic characteristic of patients in phase 1 (IQR: interquartile rate, NA: not applicable, E: extension, 5-ASA: 5-aminosalicylic acid, IMM: immunomodulator, TNF: tumor necrosis factor)

	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%)

Table 2: demographic characteristic of patients in phase 2 (IQR: interquartile rate, NA: not applicable,
E: extension, 5-ASA: 5-aminosalicylic acid, IMM: immunomodulator, TNF: tumor necrosis factor)

	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%)