Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data

Ther Adv Gastrointest Endosc

2021, Vol. 14: 1–15

DOI: 10.1177/
2631774521990623

Benjamin Gutierrez Becker*, Filippo Arcadu*, Andreas Thalhammer, Citlalli Gamez Serna, Owen Feehan, Faye Drawnel, Young S. Oh and Marco Prunotto

Abstract

Introduction: The Mayo Clinic Endoscopic Subscore is a commonly used grading system to assess the severity of ulcerative colitis. Correctly grading colonoscopies using the Mayo Clinic Endoscopic Subscore is a challenging task, with suboptimal rates of interrater and intrarater variability observed even among experienced and sufficiently trained experts. In recent years, several machine learning algorithms have been proposed in an effort to improve the standardization and reproducibility of Mayo Clinic Endoscopic Subscore grading.

Methods: Here we propose an end-to-end fully automated system based on deep learning to predict a binary version of the Mayo Clinic Endoscopic Subscore directly from raw colonoscopy videos. Differently from previous studies, the proposed method mimics the assessment done in practice by a gastroenterologist, that is, traversing the whole colonoscopy video, identifying visually informative regions and computing an overall Mayo Clinic Endoscopic Subscore. The proposed deep learning–based system has been trained and deployed on raw colonoscopies using Mayo Clinic Endoscopic Subscore ground truth provided only at the colon section level, without manually selecting frames driving the severity scoring of ulcerative colitis.

Results and Conclusion: Our evaluation on 1672 endoscopic videos obtained from a multisite data set obtained from the etrolizumab Phase II Eucalyptus and Phase III Hickory and Laurel clinical trials, show that our proposed methodology can grade endoscopic videos with a high degree of accuracy and robustness (Area Under the Receiver Operating Characteristic Curve $= 0.84$ for Mayo Clinic Endoscopic Subscore $\geqslant 1$, 0.85 for Mayo Clinic Endoscopic Subscore $\geqslant 2$ and 0.85 for Mayo Clinic Endoscopic Subscore $\geqslant 3$) and reduced amounts of manual annotation.

Plain language summary

Patient, caregiver and provider thoughts on educational materials about prescribing and medication safety

Artificial intelligence can be used to automatically assess full endoscopic videos and estimate the severity of ulcerative colitis. In this work, we present an artificial intelligence algorithm for the automatic grading of ulcerative colitis in full endoscopic videos. Our artificial intelligence models were trained and evaluated on a large and diverse set of colonoscopy videos obtained from concluded clinical trials. We demonstrate not only that artificial intelligence is able to accurately grade full endoscopic videos, but also that using diverse data sets obtained from multiple sites is critical to train robust AI models that could potentially be deployed on real-world data.

Correspondence to:
Young S. Oh
Product Development, Genentech, Inc., 1 DNA Way, South San Francisco, CA, USA.
ohy3@gene.com

Marco Prunotto
School of Pharmaceutical Sciences, Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Rue Michel-Servet 1, 1211 Geneva 4, Switzerland.
Immunology, Infectious Disease & Ophthalmology, Roche, Basel, Switzerland
marco.prunotto@unige.ch

Benjamin Gutierrez Becker
Filippo Arcadu
Andreas Thalhammer
Citlalli Gamez Serna
Owen Feehan
Roche Pharma Research and Early Development Informatics, Roche Innovation Center Basel, Basel, Switzerland

Faye Drawnel
Roche Personalized Healthcare, Genentech, Inc., South San Francisco, CA, USA

Product Development, Genentech, Inc., South San Francisco, CA, USA

*Benjamin Gutierrez Becker and Filippo Arcadu shared first authorship.

Keywords: artificial intelligence, deep learning, endoscopy, inflammatory bowel disease, ulcerative colitis

Received: 9 November 2020; revised manuscript accepted: 3 January 2021.

Introduction

Ulcerative colitis (UC) is an inflammatory bowel disease (IBD) that affects the colon mucosa and characterized by inflammation and the presence of ulcers in the large intestine and in the rectum.[1] One of the essential components to assess the status of UC patients as well as their response to therapy is the endoscopic assessment of mucosal healing. Endoscopic healing has shown to predict steroid-free and sustained clinical remission, making it a pivotal endpoint in clinical trials.[2,3]

Several scoring systems have been developed to capture the multiple findings of endoscopic examinations.[4–13] Among them, the Mayo Clinic Endoscopic Subscore (MCES),[14] a component of the overall Mayo Score[15] is the most commonly used. The MCES classifies the severity of mucosal damage in four different categories: normal or inactive, mild, moderate or severe disease. Scoring an endoscopic video using the MCES is labor intensive and is often time-critical in clinical trials. In addition, despite the widespread use of MCES in clinical trials to assess disease severity and therapeutic response in UC patients,[16,17] the MCES still relies on the subjective interpretation of colonoscopy videos and has not been fully validated.[1]

Reliable and reproducible MCES grading still represents a major limitation when monitoring UC patients as it poses a non-trivial challenge even to experienced and sufficiently trained experts.[1,18] Artificial intelligence (AI) has the potential to circumvent these limitations by providing higher throughput, and more standardized and easy-to-access scoring systems. Independent studies have explored the feasibility of developing AI systems to assess the severity of UC from still images.[19–21]

The recent publication by Yao and colleagues[22] and the concurrent work to ours by Gottlieb and colleagues[23] evaluated the use of an MCES scoring system for unaltered endoscopic videos obtained from clinical trials. These studies demonstrated not only that evaluating full videos poses a greater challenge when compared to the grading of still frames but also shown that a significant drop in performance is to be expected when a model is evaluated on an external set of colonoscopies obtained from a multicenter study as part of a clinical trial.

In this work, we present an end-to-end computer-assisted diagnosis (CAD) system based on deep learning (DL) for the automatic assessment of the MCES in high-definition white light endoscopy videos. However, previous approaches used time-consuming labor-intensive per-frame annotations, our system has been trained on colonoscopy videos with associated annotations at the level of colon sections. Despite the "weak" nature of the available ground truth—it is unknown which frame drives the score of the video—we show that it is possible to build reliable DL CAD systems able to automatically score frames in terms of the UC severity. Moreover, our model was able to perform automatic scoring of raw colonoscopy videos, without the need of pre-selecting clinically meaningful individual frames. In addition, we present the evaluation of our model in a multicenter data set obtained from clinical trials.

Methods

Data set

Our models are trained and validated on sigmoidoscopy videos selected from the Eucalyptus (ClinicalTrials.gov, NCT01336465), Hickory (ClinicalTrials.gov, NCT02100696), and Laurel (ClinicalTrials.gov, NCT02165215) Genentech/Roche clinical trial studies to test etrolizumab in patients with moderate to severe UC.

Etrolizumab is a dual-action anti-integrin antibody designed to selectively target a specific part (β7 subunit) of two key proteins, α4β7 and αEβ7 integrins, found on cells that play a key role in inflammation in IBD.

Eucalyptus is a randomized, double-blind, placebo-controlled, multicenter phase-II study involving 124 participants, 24 sites, and concluded in 2012 to evaluate the efficacy and safety of etrolizumab.[24] Hickory and Laurel are recently concluded randomized, double-blind, placebo-controlled, phase-III studies, involving 609 and 359 patients with UC, respectively. Hickory evaluated the safety, efficacy, and tolerability of etrolizumab during induction (week 14) and maintenance (week 66) compared with placebo in patients with moderate-to-severe UC who have been previously exposed to anti-tumor necrosis factor α therapy. Laurel evaluated the safety and efficacy of etrolizumab compared with placebo during maintenance (week 62) among patients who were clinical responders to etrolizumab during induction (week 10), and who have moderate-to-severe UC and were anti-tumor necrosis factor α therapy naive.

Patients in Eucalyptus, Laurel, and Hickory all had moderate to severe UC. For Eucalyptus, this was defined as a Mayo Clinic Score of 5–12 (6–12 in the United States), stool frequency subscore ⩾1, and centrally read Mayo endoscopy subscore of 2–3. In Laurel and Hickory, moderate to severe disease was defined by a Mayo Clinic Score of 6–12, centrally read Mayo endoscopy subscore of 2–3, rectal bleeding subscore ⩾1, and stool frequency subscore ⩾1. In addition, documentation of colonic involvement of UC extending a minimum of 20 cm (Laurel, Hickory) or 25 cm (Eucalpytus) from the anal verge was required. Patients in Laurel were naive to TNF inhibitors but must have had an inadequate response, loss of response, or intolerance to prior immunosuppressant and/or corticosteroid treatment within 5 years of study entry. Patients in Hickory had a history of exposure to at least one TNF inhibitor within 5 years of study entry.

The colonoscopies of Hickory and Laurel are equipped with manual annotations of the colon subsection (rectum, sigmoid colon, descending colon) and the associated MCES, as assessed by the onsite investigator and reviewed by central readers.

The videos used for the training and validation of MCES prediction models (obtained from the Hickory and Laurel clinical trials) correspond to 286 different sites distributed across 28 countries.

In addition to the data obtained from the clinical trials, we externally validated the performance of our models using still frames obtained from the publicly available Hyperkvasir data set,[25] which are equipped with an MCES score assigned by the authors of the data set.

Overview of the end-to-end UC scoring model on raw colonoscopy videos

Most of the previous approaches deploying DL models for the automatic assessment of UC[21,20] involve following the procedure shown in the flow diagram in Figure 1.

1. Colonoscopy videos are obtained from a single site, and are assessed by gastroenterologists who identify frames driving the MCES score and assigns a MCES to them. This is an intensive and repetitive task that can be performed only by a trained gastroenterologist with expertise on UC.
2. A DL model is trained using these manually extracted frames together with their corresponding MCES. The trained DL model is deployed to perform predictions on still images obtained from colonoscopy videos previously unseen by the model. The DL model requires the manual selection of frames to perform the prediction. This selection has to mimic the one performed during training, and therefore involves the supervision of a trained gastroenterologist.
3. The manually selected frames are fed into the DL model, and a predicted score indicating the UC severity is obtained for each frame.

One of the most labor-intensive and difficult tasks of this approach is the manual selection of frames for the training and deployment of the DL model. Our proposed approach aims at overcoming this limitation by performing an automatic selection of *readable* frames, which can in turn be used to train a DL model for the automatic assessment of UC severity. Our proposed approach, summarized in the flow diagram is shown in Figure 2.

1. Colonoscopy videos are assessed by gastroenterologists, who assign a severity score to each colon subsection. This assessment is performed as part of the clinical trials used in this study; therefore, it does not involve a tailor-made annotation procedure for the training of a DL model.
2. We extracted frames at random from colonoscopy videos and classified them as either readable or non-readable. These annotations do not involve a clinical assessment of the frames, and therefore were performed by non-gastroenterologists.
3. A quality control (QC) model is trained using these frames. The purpose of this model is to discriminate between *readable* and *non-readable* frames. From here on, we will refer to this model as the QC model.
4. We used the QC model to automatically extract readable frames from colonoscopy

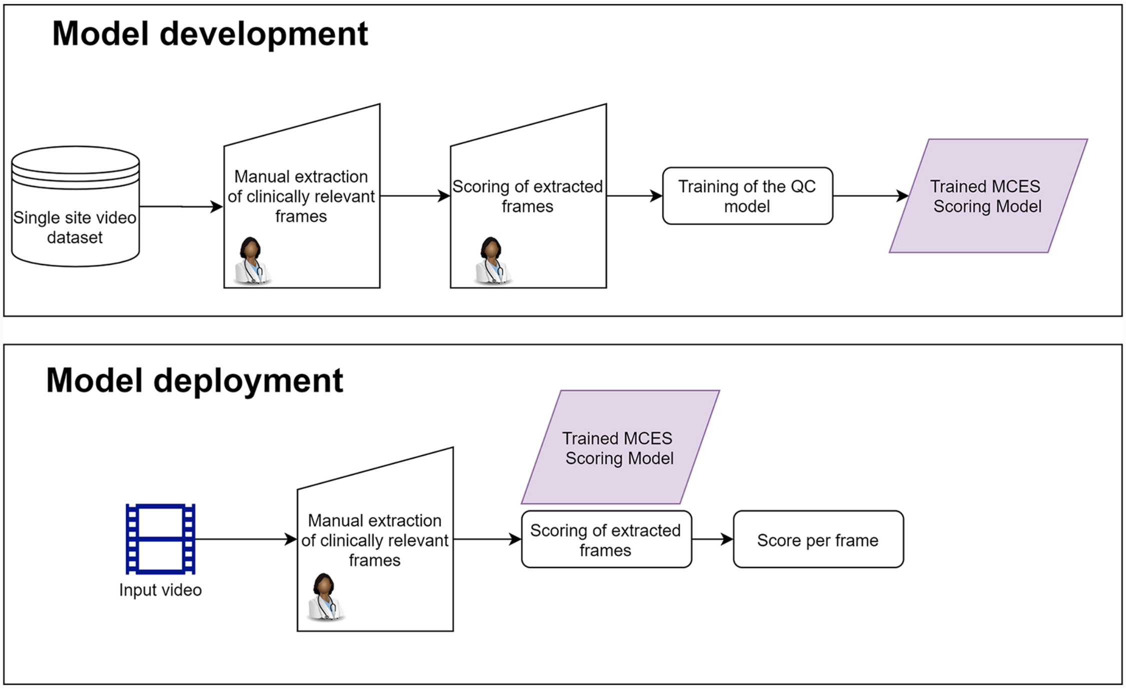

Model development
Single site video dataset
Manual extraction of clinically relevant frames
Scoring of extracted frames
Training of the QC model
Trained MCES Scoring Model
Model deployment
Input video
Manual extraction of clinically relevant frames
Trained MCES Scoring Model
Scoring of extracted frames
Score per frame

Figure 1. Flow diagram showing the standard method to perform automatic grading of ulcerative colitis on endoscopy videos.
A critical component of this approach is the careful selection of clinically relevant frames by an experienced gastroenterologist, not only during the development of the model but also during deployment.

videos. Each *readable frame* was assigned an MCES score obtained as part of the central reading of the clinical trial. Notice that this MCES score correspond to the whole colon section, and not to a specific frame.

5. The extracted readable frames, paired with their corresponding MCES score, are used to train a second DL model. This model, which we refer as the Ulcerative Colitis Scoring (UCS) model was used to automatically predict the severity of UC from still frames. The MCES score is obtained from the central readings of the clinical trial and are assigned to the correct frame through the automatic text detection described in the pre-processing section.
6. Finally, the two trained models: QC and UCS, were deployed to perform predictions on previously unseen videos. These predictions were performed in a fully automatic fashion on raw colonoscopy videos. Raw colonoscopy videos were fed to the QC model, which automatically extracts *readable* frames; the UCS model is then used to obtain a severity score for each readable frame. A final score for an entire colon section is obtained by aggregating the individual scores of each frame.

QC model for the automatic extraction of readable frames

One of the major challenges when developing AI systems operating directly on full colonoscopy videos is that a large proportion of the frames within the video do not allow for an appropriate visual assessment of the mucosa. Several factors can contribute to render a frame *non-readable* such as: presence of water and bubbles, blurriness, presence of stool, pixel *saturation*, the camera being too close to the mucosa, or a combination of these factors.[26] Identifying and extracting readable frames is key to construct a meaningful training set for the DL model targeting the prediction of the MCES score.

The first stage of our system for MCES prediction was a QC model. This model summarizes an entire endoscopic video into a limited subset of readable frames.

Training of the QC model. The main component of the developed QC model is a Convolutional Neural Network (CNN) trained to categorize frames in a video between two classes: *readable* and *unreadable* frames. We trained the network using 5000 still images extracted randomly from 351 videos from the Eucalyptus clinical trial. It is

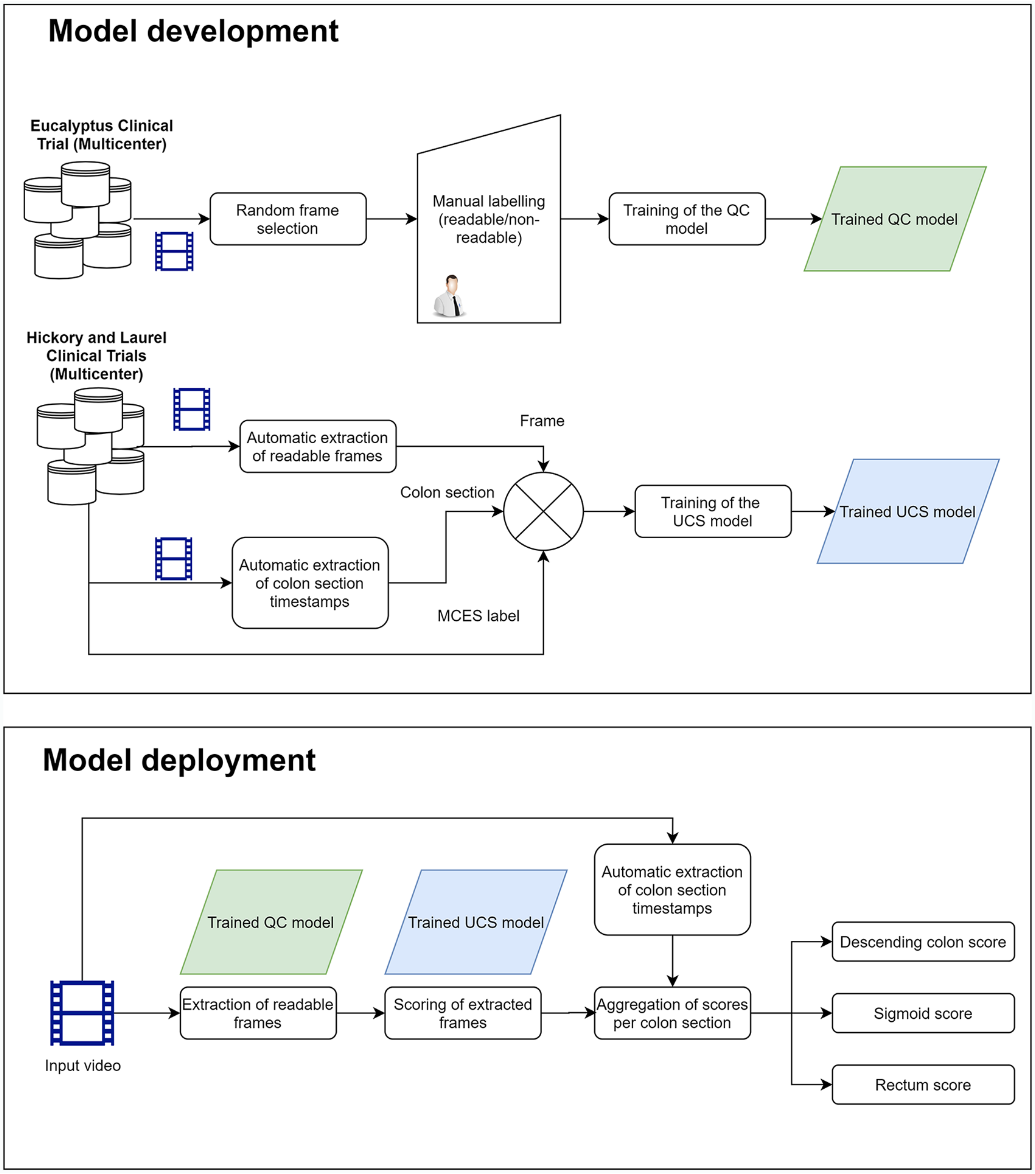

Model development
Eucalyptus Clinical Trial (Multicenter)
Random frame selection
Manual labelling (readable/non-readable)
Training of the QC model
Trained QC model
Hickory and Laurel Clinical Trials (Multicenter)
Automatic extraction of readable frames
Frame
Colon section
Automatic extraction of colon section timestamps
MCES label
Training of the UCS model
Trained UCS model
Model deployment
Trained QC model
Trained UCS model
Automatic extraction of colon section timestamps
Input video
Extraction of readable frames
Scoring of extracted frames
Aggregation of scores per colon section
Descending colon score
Sigmoid score
Rectum score

Figure 2. Flow diagram showing the procedure for the development and the deployment of the proposed model. Different to previous approaches, our model is based on readings performed as part of the clinical trials, and on quality annotations performed by non-expert. Therefore, our proposed methodology does not require ad hoc annotations by expert gastroenterologists neither during training nor on the deployment of the model.

important to mention that the videos from the Eucalyptus trial were used *only* to train and validate the informative frame extraction algorithm. These frames were not used in the development of the MCES scoring system itself.

Two annotators labeled manually each frame in one of the two classes: *readable* or unreadable. The following set of rules was used to determine when a video frame should be considered informative: (1) the colon walls are in the Field of View (FOV), (2) contrast and sharpness of the frame are sufficient to visually inspect, at least partially, the mucosa and its vascular pattern, (3) there is no presence of artifacts obstructing completely the visibility of the mucosa within the FOV (stool, water, bubbles, reflections, etc.). If a frame fulfilled all these criteria, it was considered *readable*; otherwise the frame was deemed *unreadable*.

The two annotators achieved a very high inter-rater agreement (Cohen's kappa = 0.91) in this task. Since the aforementioned criteria are based

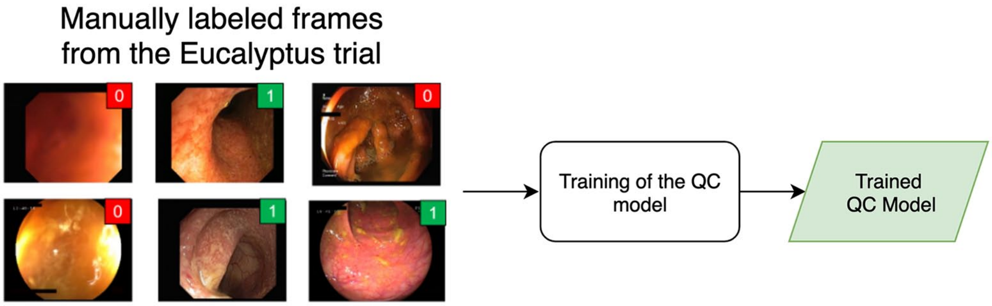


Figure 3. Overview of the training of the Quality Control (QC) model.
The QC model is trained using still frames obtained from the Eucalyptus clinical trial. Frames are manually annotated as readable or non-readable by two annotators. The output of the model is a quality score reflecting how readable a frame is.

only on the overall visual quality of the frames and they do not involve any clinical evaluation, none of the annotators was a gastroenterologist. This means that we were able to train the QC model (Figure 3) without the need of supervision by expert gastroenterologists.

Frame sampling algorithm. The second element of the QC model is a frame-sampling algorithm. This algorithm uses the QC scores provided by the CNN to automatically extract readable frames from a complete colonoscopy video. The proposed frame extraction algorithm consists of the following steps: (1) the full colonoscopy video is divided into non-overlapping windows of 5 s duration; (2) all frames within a selected window are equipped with a quality score provided by the QC model. The frame with the highest quality score in the window is selected; (3) if the quality score of the selected frame is above a predetermined threshold (0.99 in our experiments), the frame is kept. Otherwise, the frame and the entire corresponding window are discarded. This sampling algorithm described above allowed us to summarize the complete endoscopic video into a small subset of high visual quality frames covering the entire length of the videos.

UCS model for the automatic grading of still frames

The next component of our system is a model developed to automatically grade UC on colonoscopy videos. This model consisted of two components: a CNN trained to automatically score the severity of UC on still frames, given their visual appearance, and an aggregation algorithm, which allowed our model to summarize the individual predictions of each individual frame into a single score for an entire colon section.

Training of the CNN for automatic scoring of UC. We trained a model for the automatic estimation of UC. This model was trained in a similar fashion to previous approaches that model automatic scoring of UC as a binary classification task.[21,20] The fundamental difference between our model and previous efforts is that our model was not trained using frames that were carefully selected and graded by an expert gastroenterologists. Instead, we used frames that were automatically extracted using the QC model described above, and we assigned an MCES *weak label* to them. These labels are obtained automatically by getting a hold of the colon section associated with each frame using a text detection algorithm and by pairing it with the MCES grade obtained from the clinical trials (see Apendix A for a detailed description of the pre-processing stage). These readings are not associated with a specific frame driving the MCES diagnosis, but rather to the subset of the video covering a colon segment, thus they represent weak labels. The scoring model is trained as a binary classifier with the task to discriminate between frames corresponding to an MCES below or above a pre-specified threshold.

Deployment of the end-to-end scoring system on full endoscopic videos

After training the individual components of the end-to-end scoring system (the QC model and the UCS model), we evaluated the MCES in a set of videos that were not used for training the system. The scoring of an endoscopic video (Figure 4) is performed as follows: (1) a subset of informative frames is sampled from the video using the QC model, (2) the estimated UC severity for each frame is assessed using the UCS model, (3) the full MCES of a colon subsection

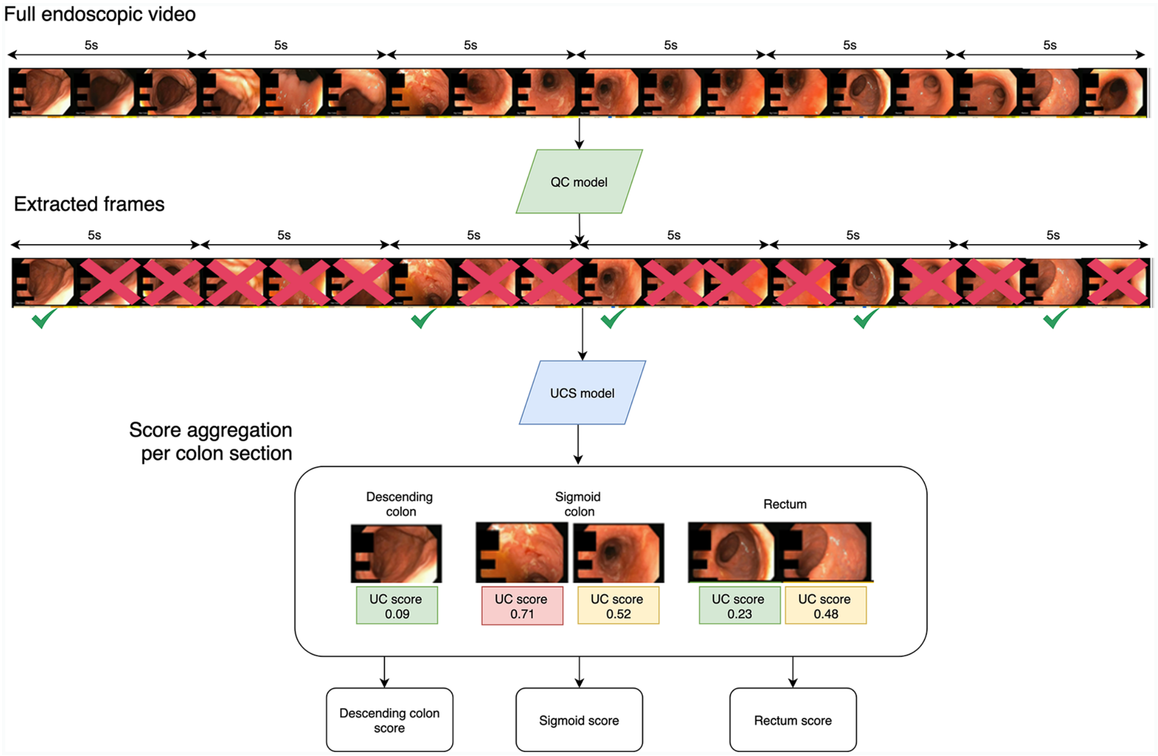


Figure 4. Overview of the end-to-end MCES scoring on full endoscopic videos.
First, readable frames are extracted from the video using the QC model. Each individual frame is passed through the UCS model to obtain an MCES score per-frame. The individual scores are aggregated and summarized into a single MCES per colon section.

is obtained by aggregating the individual scores of each frame. This aggregation is performed by averaging the scores of all frames belonging to a colon subsection.

Results

Performance of the QC model

We evaluated the performance of the proposed QC model on the task of discriminating between readable and non-readable frames. This evaluation was performed using a total of 5000 frames obtained from Eucalyptus that were manually labeled by two non-expert readers. Those frames for which the readers did not agree were removed from the analysis, leaving 4371 frames to train and evaluate the QC model. The remaining frames were split in 5 folds, each one with its own training, tuning, and external validation sets (70%/20%/10%, see Appendix 2 for a full description of the splitting strategy). Our proposed QC model shows an excellent performance as demonstrated by the receiver operating characteristic (ROC) curve in Figure 5 with an AUROC of 0.98 ± 0.0022.

Evaluation of the UCS prediction algorithm on raw videos

Evaluation of the MCES prediction algorithm was performed on raw colonoscopy videos obtained from the Hickory and Laurel clinical trials. A total of 1672 sigmoidoscopy videos obtained from 286 different sites and corresponding to 1105 patients were selected for analysis. Data were split in 5-folds following the strategy described in Appendix 3. The UCS model was trained on multiple binary tasks, each one discriminating above a determined threshold (MCES $\geqslant$ 1, MCES $\geqslant$ 2 or MCES $\geqslant$ 3). The models we trained obtained an AUROC of 0.84 ± 0.0237 for MCES $\geqslant$ 1 with a (precision 0.92 ± 0.021, recall 0.79 ± 0.060), an AUROC of 0.85 ± 0.0222 for MCES $\geqslant$ 2 (precision 0.85 ± 0.048, recall 0.81 ± 0.070) and 0.85 ± 0.0099 for MCES $\geqslant$ 3 (precision 0.81 ± 0.075, recall 0.77 ± 0.050). The ROC curves for these experiments are presented in Figure 6.

Evaluation of the UCS model on still frames obtained from an external data set

We also evaluated the performance of our UCS model on the task of scoring individual still

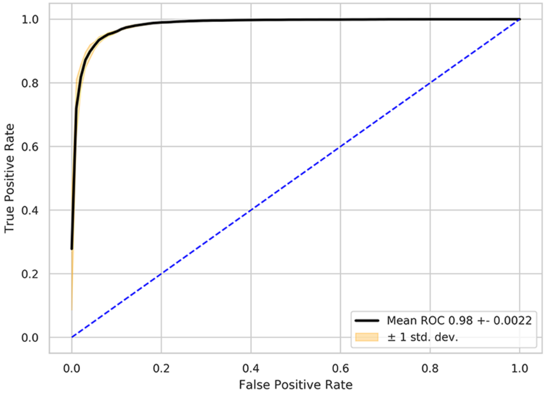


Figure 5. Receiver operating curve (ROC) showing the performance obtained on the binary task of discriminating between readable and non-readable frames obtained from videos of the Eucalyptus clinical trial.

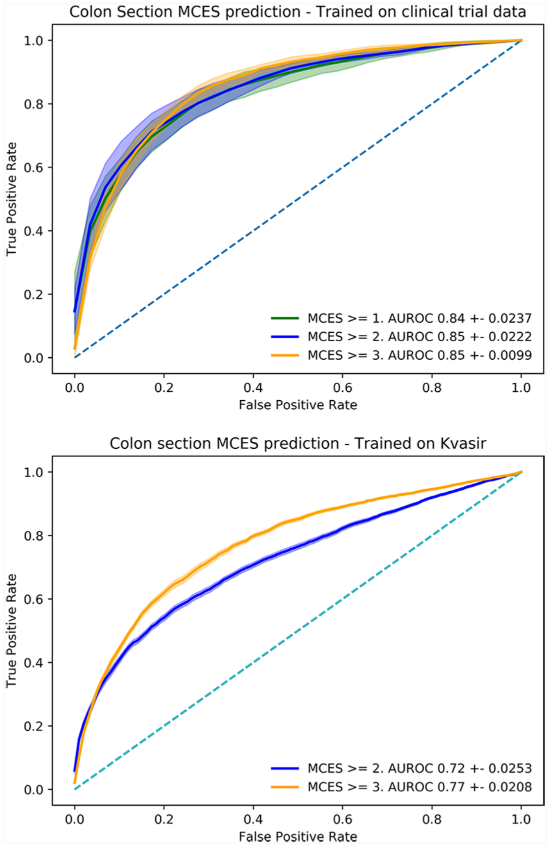


Figure 6. Receiver operating curve (ROC) obtained for the evaluation of models trained for the automatic scoring of MCES of full videos obtained from the Hickory and Laurel clinical trials. Top: performance of the models trained on the Hickory and Laurel clinical trials. Bottom: performance of the models trained using frames obtained from the HyperKvasir data set.

frames. The evaluation of our models on still frames was performed using the publicly available HyperKvasir data set. The HyperKvasir data set contains a set of labeled still frames corresponding to MCES scores of 1 (201 frames), 2 (444 frames), and 3 (133 frames). For the evaluation on individual still frames, we used the UCS model trained using raw endoscopic videos from clinical trials, but since the frames of the HyperKvasir data s*et al*ready contains carefully selected endoscopic frames the QC component of our model was not used for this experiment. Our UCS model trained on raw colonoscopy videos obtained AUROCs of 0.82 ± 0.0212 for MCES ⩾ 2 (precision 0.92 ± 0.024, recall 0.73 ± 0.78) and 0.83 ± 0.0395 (precision 0.39 ± 0.061, recall 0.84 ± 0.060) for MCES ⩾ 3. As a comparison, we trained and evaluated CNN models using the official splits of the HyperKvasir data set. These models are trained and evaluated *only* on the HyperKvasir data set and did not include the data obtained from the clinical trials. These models were trained using a standard CNN architecture (Resnet50) in a similar fashion as previous approaches operating on single frames.[20,21] The models trained and evaluated on HyperKvasir obtained an AUROC of 0.85 ± 0.0273 and 0.91 ± 0.0398 for MCES ⩾ 2 and MCES ⩾ 3, respectively. The ROC curves for these experiments are plotted on Figure 7.

Finally, the models trained on single frames obtained from HyperKvasir were evaluated on the full raw endoscopic videos obtained from the clinical trials. The HyperKvasir models obtained an AUROC of 0.72 ± 0.0253 for MCES ⩾ 2 and 0.77 ± 0.0208 for MCES ⩾ 3.

Discussion

In this work, we have presented an end-to-end DL-based system to automatically assess the severity of UC from endoscopic videos. Differently from previous efforts using AI to grade MCES, our approach does not leverage high-quality clinically meaningful frames, carefully extracted and annotated by medical experts. Instead, it directly operates on raw endoscopic videos, which are automatically pre-processed, screened for visual quality, and finally fed to a CNN for training. The proposed methodology is characterized by its accurate and robust performance, as shown in the presented experiments. Our study was evaluated on a large and diverse data set that was collected

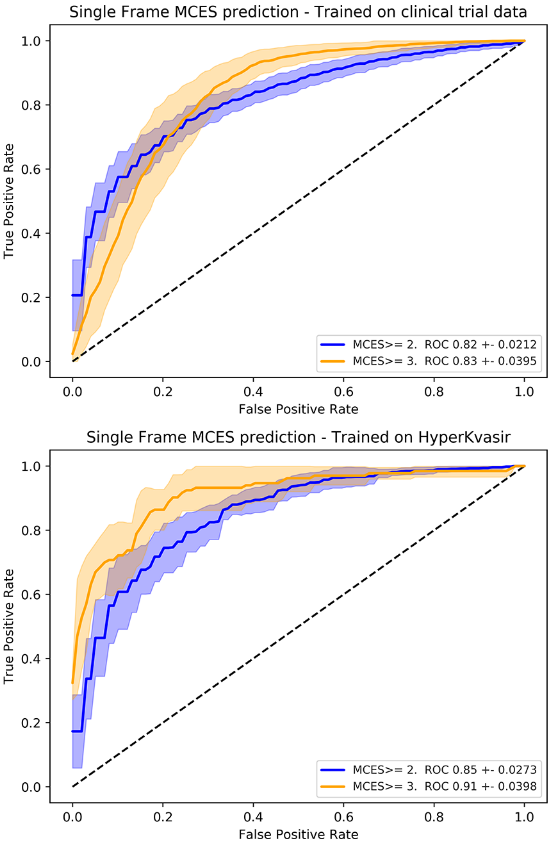


Figure 7. Receiver operating curve (ROC) obtained for the evaluation of models trained for the automatic scoring of MCES on individual frames obtained from the HyperKvasir data set. Top: performance of the models trained on the Hickory and Laurel clinical trials. Bottom: performance of the models trained using frames obtained from the HyperKvasir data set.

and annotated at multiple sites and obtained from different countries.

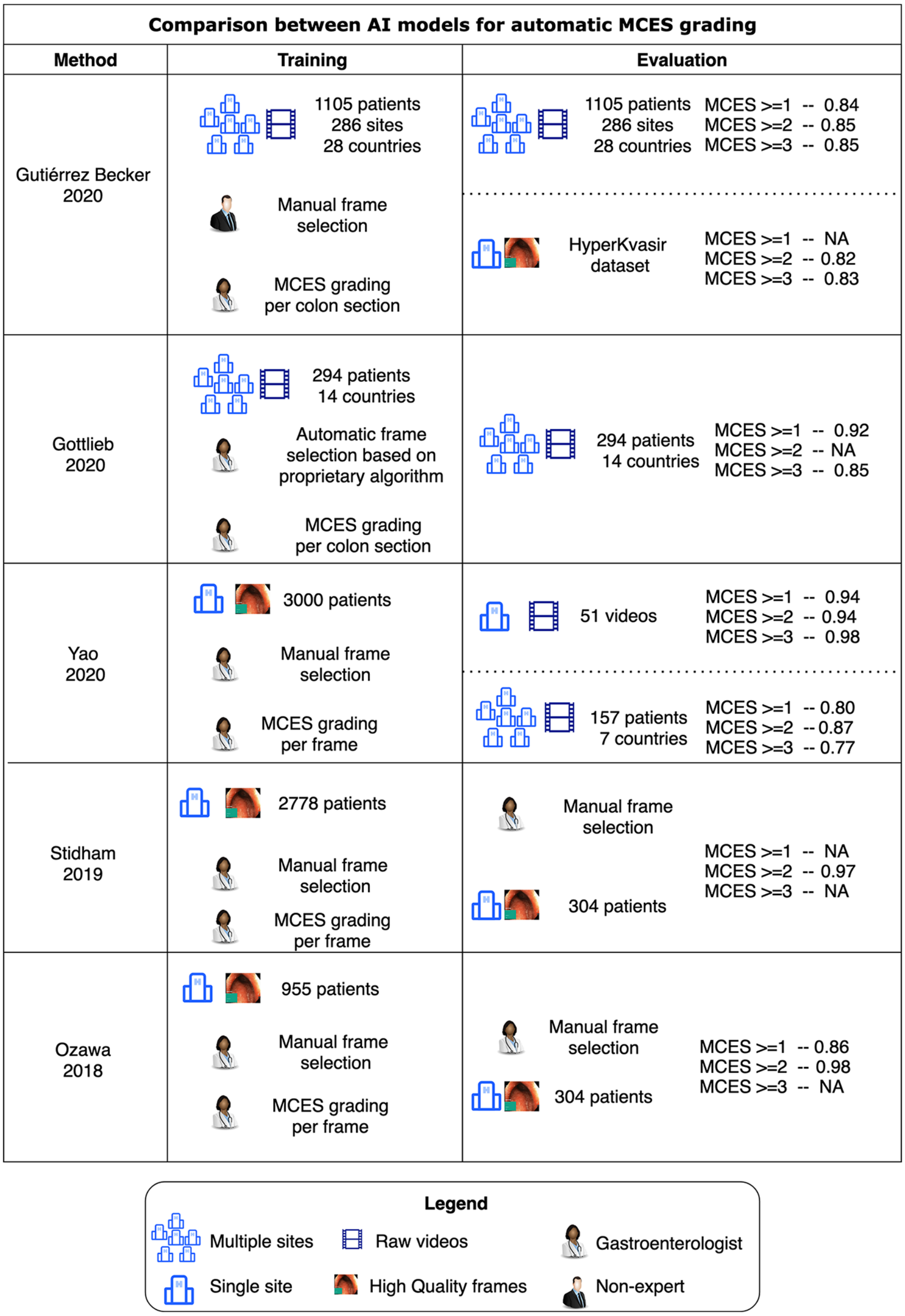
Availability of large databases of annotated images is a major bottleneck when engineering AI-based diagnostic tools for medical imaging. This challenge is particularly evident when developing end-to-end systems for endoscopic image analysis: finding informative and clinically significant frames is a challenging problem by itself and incredibly time-consuming. Our study shows that devising an end-to-end approach operating on full videos is a feasible alternative solution that does not require the strenuous, tedious, and time-consuming task of manually selecting and scoring individual frames, which can only be performed by expert gastroenterologists.

Developing end-to-end systems operating directly on endoscopic videos is a substantially more challenging task compared to performing predictions on single images or still frames. The main challenge when operating directly on raw videos is to pinpoint the clinically informative frames within each video, as it has been reported by multiple previous works in the field.[27–29]

In Figure 8, we present a comparison between our approach and the previously published studies on automatic MCES prediction by Ozawa and colleagues,[21] Stidham and colleagues,[20] Yao and colleagues,[22] and the concurrent work of Gottlieb and colleagues.[23] Our motivation is to enable a fair assessment of our approach in comparison to previous efforts, but it is important to note that the experimental conditions between these studies are remarkably different.

The first major difference between our work and the aforementioned models for automatic MCES prediction is the nature of the training data sets. Previous models were trained using carefully curated data sets, where clinically representative frames of each MCES score were extracted and annotated by experts. These frames are manually selected as representative examples of each MCES grade, a tedious time-consuming annotation task performed by experienced gastroenterologists. We have proposed instead to use weak labels automatically obtained from a clinical trial. Our experiments show that this is a reasonable alternative to train an AI-based system for MCES grading. Weak labels contain extremely valuable information that can be leveraged when training AI models and although weak labels cannot be as informative as carefully curated and annotated still frames, our experiments suggest that this limitation is compensated by the amount of images and the diversity in appearance that can be leveraged when using a weakly supervised approach.

The second important difference highlighted in Figure 8 between our work and previous efforts is the fact that our study was conducted on a highly heterogeneous set of endoscopic videos, collected from hundreds of different sites with various devices and assessed by hundreds of investigators. Conducting our experiments in such a diverse setting, allows us to test our models in a setting closer to the real-world application. In Figure 8, we can observe that the diversity of the data set used for the training and evaluation of MCES prediction models has a remarkable impact in the final performance of this models. Each of the models trained using our weak labeling strategy were trained using an average of ~200,000 frames



Comparison between AI models for automatic MCES grading		
Method	**Training**	**Evaluation**
Gutiérrez Becker 2020	1105 patients 286 sites 28 countries Manual frame selection MCES grading per colon section	1105 patients 286 sites 28 countries MCES >=1 -- 0.84 MCES >=2 -- 0.85 MCES >=3 -- 0.85 HyperKvasir dataset MCES >=1 -- NA MCES >=2 -- 0.82 MCES >=3 -- 0.83
Gottlieb 2020	294 patients 14 countries Automatic frame selection based on proprietary algorithm MCES grading per colon section	294 patients 14 countries MCES >=1 -- 0.92 MCES >=2 -- NA MCES >=3 -- 0.85
Yao 2020	3000 patients Manual frame selection MCES grading per frame	51 videos MCES >=1 -- 0.94 MCES >=2 -- 0.94 MCES >=3 -- 0.98 157 patients 7 countries MCES >=1 -- 0.80 MCES >=2 -- 0.87 MCES >=3 -- 0.77
Stidham 2019	2778 patients Manual frame selection MCES grading per frame	Manual frame selection 304 patients MCES >=1 -- NA MCES >=2 -- 0.97 MCES >=3 -- NA
Ozawa 2018	955 patients Manual frame selection MCES grading per frame	Manual frame selection 304 patients MCES >=1 -- 0.86 MCES >=2 -- 0.98 MCES >=3 -- NA

Legend

Multiple sites	Raw videos	Gastroenterologist
Single site	High Quality frames	Non-expert

Figure 8. Graphical summary of the data set characteristics and the experimental results reported by Gottlieb and colleagues,[23] Ozawa and colleagues,[21] Stidham and colleagues,[20] Yao and colleagues,[22] and by this study.

which contrasts with the 778 frames used to train the models using the HyperKvasir data set.

AI-based systems for diagnosis are prone to be biased, and to exploit confounding factors when trained on data sets obtained from a limited number of sites.[30] Training and evaluating an AI-model on heterogeneous clinical trial data, makes a successful translation to daily practice more likely compared to models trained on data collected at a single site. The difference on performance between the models trained on high-resolution data obtained from single-center data sets (Stidham and colleagues, Ozawa and colleagues) *versus* those evaluated on clinical trials (Gottlieb and colleagues, Yao and colleagues, Gutierrez and colleagues), suggest that single-center studies might be over optimistic in their results and that those models might suffer from drops in performance when deployed on real-world data. This hypothesis is supported by our own experiments summarized in Figures 6 and 7 where models trained and evaluated on the HyperKvasir data set have a high performance in terms of their AUROC (0.85 for MCES ⩾ 2, 0.91 of MCES ⩾ 3). However, the exact same models resulted in considerably lower AUROCs when deployed on clinical trial data (0.72 for MCES ⩾ 2, 0.77 for MCES ⩾ 3). These results suggest that the acquisition and annotation of diverse and large data sets of colonoscopy videos is a requirement for the deployment of AI models which can be robustly applied in standard clinical settings where the appearance of endoscopic videos is diverse. Our models and evaluations were limited to the binary case, where the prediction performed by the AI model is the probability of a video being above a certain MCES score. These models could potentially be useful as a prescreening tool to identify patients above certain MCES; however, methods that provide a full ordinal score would be more useful for clinical applications.

In this work, we have focused on the problem of automated grading of endoscopic videos. Our results encourage us to believe that AI-based systems have potential to expand our capabilities to analyze endoscopic videos. AI allows a frame-by-frame analysis of videos, which can in turn lead to a more accurate assessment of disease burden. Presumably, disease burden is important for UC patients, as demonstrated by recent scoring systems for UC which aim at including information regarding the location and extent of the disease.[30,31]. Although this study focused on UC, a similar framework might be extended to Crohn's disease, whose disease burden is still evaluated in a suboptimal way by means of either the Simple Endoscopic Score for Crohn's disease (SES-CD) or the Crohn's Disease Index of Severity (CDEIS).[32]

Automation through AI can bring great benefits in terms of standardization of complex diagnosis such as UC and CD, but the vision inspiring this work goes beyond automation. AI-based diagnostic tools could greatly facilitate the early identification of IBD patients as well as speeding-up patient recruitment for clinical trials, which is key to enable shorter and cost-effective trials. Moreover, we believe that leveraging AI to extract insights from endoscopic data as well as from other data sources like histology, spatial transcriptomic, stool proteomics, etc, could substantially help to improve our understanding of the pathogenesis of UC and CD, which is still very limited and will require to combine bits of information from all these different types of data.

Conclusion

We have developed an end-to-end AI CAD system for the automatic assessment of MCES from endoscopic videos collected by multiple sites in the context of clinical trials targeting UC. The high agreement between the proposed end-to-end CAD system for MCES prediction and human reviewers show that such a tool for automatic grading can be trained without having to create time-consuming and expensive data sets consisting of frames handpicked from each video by expert gastroenterologists.

A full automation of endoscopic grading system assisted by AI has the potential to lead to faster, more repeatable and objective endoscopic assessments, which in turn can lead to more efficient and standardized diagnosis of UC in the clinical setting.

Future work will focus on further developing the models presented here by leveraging larger cohorts of endoscopy videos from UC patients. Moreover, we plan to extend the proposed approach to target the identification of the MCES components such as erythema, friability, ulcers, and spontaneous bleeding, that would pave the way for a more granular automated perspective of the UC status of a patient compared to a single number such as the MCES.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of interest statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Benjamin Gutierrez Becker https://orcid.org/0000-0002-5506-7785
Andreas Thalhammer https://orcid.org/0000-0002-0991-5771

References

1. Mohammed Vashist N, Samaan M, Mosli M, *et al*. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. *Cochrane Database Syst Rev* 2018; 1: CD011450.

2. Iacucci M, Furfaro F, Matsumoto T, *et al*. Advanced endoscopic techniques in the assessment of inflammatory bowel disease: new technology, new era. *Gut* 2019; 68: 562–572.

3. United States Food and Drug Administration. Ulcerative colitis: clinical trial endpoints guidance for industry, 2016, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/ulcerative-colitis-clinical-trial-endpoints-guidance-industry

4. Truelove S and Witts L. Cortisone in ulcerative colitis. *Br Med J* 1955; 2: 1386.

5. Baron JH, Connell AM and Lennard-Jones JE. Variation between observers in describing mucosal appearances in proctocolitis. *Br Med J* 1964; 1: 89–92.

6. Powell-Tuck J, Day DW, Buckell NA, *et al*. Correlations between defined sigmoidoscopic appearances and other measures of disease activity in ulcerative colitis. *Dig Dis Sci* 1982; 27: 533–537.

7. Sutherland LR and Martin F. 5-Aminosalicylic acid enemas in treatment of distal ulcerative colitis and proctitis in Canada. *Dig Dis Sci* 1987; 32(Suppl. 12): 64S–66S.

8. Rachmilewitz D, Barbier F, Defrance P, *et al*. Coated mesalazine (5-aminosalicylic acid) versus sulphasalazine in the treatment of active ulcerative colitis: a randomised trial. *Br Med J* 1989; 298: 82–86.

9. Feagan BG, Greenberg GR, Wild G, *et al*. Treatment of ulcerative colitis with a humanized antibody to the α4β7 integrin. *N Engl J Med* 2005; 352: 2499–2507.

10. Naganuma M, Ichikawa H, Inoue N, *et al*. Novel endoscopic activity index is useful for choosing treatment in severe active ulcerative colitis patients. *J Gastroenterol* 2010; 45: 936–943.

11. Travis SPL, Schnell D, Krzeski P, *et al*. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). *Gut* 2012; 61: 535–542.

12. Samuel S, Bruining DH, Loftus EV Jr, *et al*. Validation of the Ulcerative Colitis Colonoscopic Index of severity and its correlation with disease activity measures. *Clin Gastroenterol Hepatol* 2013; 11: 49–54.

13. Lobatón T, Bessissow T, De Hertogh G, *et al*. The Modified Mayo Endoscopic Score (MMES): a new index for the assessment of extension and severity of endoscopic activity in ulcerative colitis patients. *J Crohns Colitis* 2015; 9: 846–852.

14. Schroeder KW, Tremaine WJ and Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. *N Engl J Med* 1987; 317: 1625–1629.

15. Rutgeerts P, Sandborn WJ, Feagan BG, *et al*. Infliximab for induction and maintenance therapy for ulcerative colitis. *N Engl J Med* 2005; 353: 2462–2476.

16. Maaser C, Sturm A, Vavricka SR, *et al*. ECCO-ESGAR guideline for diagnostic assessment in IBD part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohns Colitis* 2019; 13: 144–164.

17. Rubin DT, Ananthakrishnan AN, Siegel CA, *et al*. ACG clinical guideline: ulcerative colitis in adults. *Am J Gastroenterol* 2019; 114: 384–413.

18. 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: 969–973.

19. Alammari A, Islam AR, Oh J, *et al*. *Classification of ulcerative colitis severity in colonoscopy videos using CNN*. In: *Proceedings of the 9th international conference on information management and engineering*, Barcelona, October 2017, pp. 139–144. New York: ACM.

20. 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.

21. 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: 416–421.

22. Yao H, Najarian K, Gryak J, *et al*. Fully automated endoscopic disease activity assessment in ulcerative colitis. *Gastrointest Endosc*. Epub ahead of print 15 August 2020. DOI: 10.1016/j.gie.2020.08.011

23. Gottlieb K, Requa J, Karnes W, *et al*. Central reading of ulcerative colitis clinical trial videos

using neural networks. *Gastroenterology*. Epub ahead of print 21 October 2020. DOI: 10.1053/j.gastro.2020.10.024

24. Vermeire S, O'Byrne S, Keir M, *et al.* Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial. *Lancet* 2014; 384: 309–318.
25. Borgli H, Thambawita V, Smedsrud P, *et al.* HyperKvasir: a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data* 2020; 7: 1–14.
26. Ali S, Zhou F, Braden B, *et al.* An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci Rep* 2020; 10: 1–34.
27. Atasoy S, Mateus D, Lallemand J, *et al.* Endoscopic video manifolds. In: Jiang T, Navab N, Pluim JPW, *et al.* (eds) *Medical image computing and computer—assisted intervention—MICCAI 2010* (Lecture notes in computer science). Vol. 6362. Berlin: Springer, 2010, pp. 437–445.
28. Bashar MK, Kitasaka T, Suenaga Y, *et al.* Automatic detection of informative frames from wireless capsule endoscopy images. *Med Image Anal* 2010; 14: 449–470.
29. Rezbaul Islam ABM, Alammari A, Oh JH, *et al.* Non-informative frame classification in colonoscopy videos using CNNs. In: *Proceedings of the 2018 3rd international conference on biomedical imaging, signal processing*, Bari, October 2018.
30. Bálint A, Farkas K, Szepes Z, *et al.* How disease extent can be included in the endoscopic activity index of ulcerative colitis: the panMayo score, a promising scoring system. *BMC Gastroenterol* 2018; 18: 7.
31. Hosoe N, Nakano M, Takeuchi K, *et al.* Establishment of a novel scoring system for colon capsule endoscopy to assess the severity of ulcerative colitis-capsule scoring of ulcerative colitis. *Inflamm Bowel Dis* 2018; 24: 2641–2647.
32. Papay P, Ignjatovic A, Karmiris K, *et al.* Optimising monitoring in the management of Crohn's disease: a physician's perspective. *J Crohns Colitis* 2013; 7: 653–669.
33. Zhou X, Yao C, Wen H, *et al.* *EAST: an efficient and accurate scene text detector*. In: *Proceedings of the 30th IEEE conference on computer vision and pattern recognition, CVPR 2017*, Honolulu, HI, 21–26 July 2017.
34. He K, Zhang X, Ren S, *et al.* *Deep residual learning for image recognition*. In: *Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition*, Las Vegas, NV, 27–30 June 2016.
35. Yosinski J, Clune J, Bengio Y, *et al.* How transferable are features in deep neural networks? In: *Advances in neural information processing systems*, 2014, https://arxiv.org/pdf/1411.1792.pdf
36. Russakovsky O, Deng J, Su H, *et al.* ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; 115: 211–252.
37. Kingma DP and Ba JL. *Adam: a method for stochastic optimization*. In: *3rd international conference on learning representations, ICLR 2015—conference track proceedings*, San Diego, CA, 7–9 May 2015.

Appendix 1

Video pre-processing and weak label extraction

As a first step, a suite of pre-processing algorithms (Figure 9) was applied to the entire colonoscopy data set obtained from the Eucalyptus, Hickory, and Laurel clinical trials.

1. The FOV and imprinted text within each frame were identified. Each frame was subsequently processed to mask out the text in the image. An FOV mask was created by applying minimum thresholding to the average frame per video and its standard deviation, followed by the union of both obtained masks. Undesired imprinted text on frames was masked out, after identification with the Efficient and Accurate Scene Text Detector (EAST) CNN.[33]
2. Images were cropped to remove regions outside the FOV. The cropped image was resized to the standard size expected by the architecture of the CNNs (224 × 224 × 3 pixels).
3. Each video belonging to the Hickory and Laurel trials contained time stamps corresponding to the start and end point of each colon section within the video. This information was encoded in the form of text appearing within the frame. We used these time stamps to automatically extract the corresponding colon section for each frame of the video and to establish a link to the colon-section-wise MCES provided by the central readers. These assignments of MCES scores to larger video segments (one per colon section) constitute *weak labels* that are further used to train the UC grading algorithm.

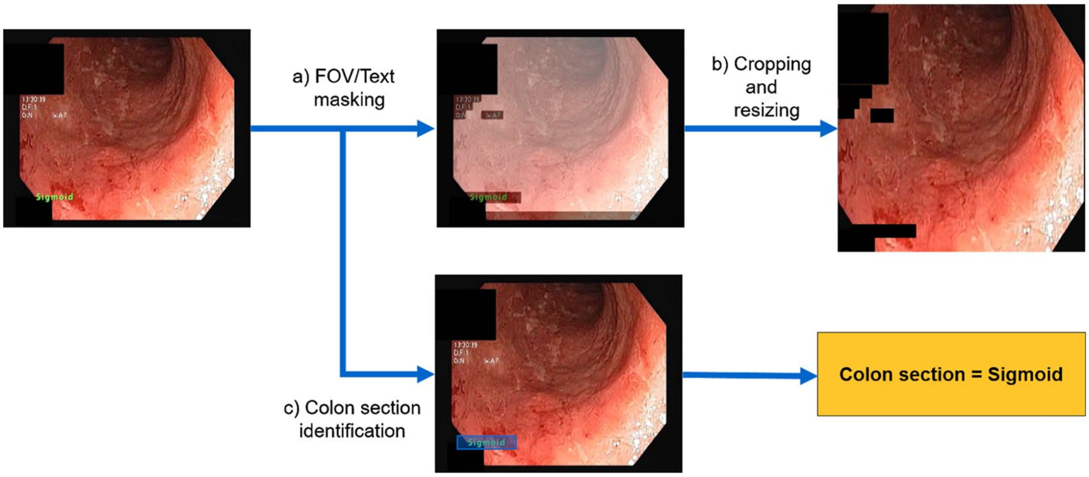


Figure 9. A set of pre-processing steps are applied to each video in the dataset: (a) field of view (FOV) and text masks are obtained for each frame, (b) each frame is cropped to remove areas outside the FOV, and undesired imprinted text is masked out, and (c) the text identifying the current colon section is identified, and extracted to obtain a weak label associated to the frame.

Appendix 2

Implementation details

The framework was implemented using Python 3.6, Pytorch 1.16, OpenCV 4.01, and Pytorch Lightning 0.9.1. All DL models were trained for using a ResNet50 architecture,[34] by transfer learning[35] from ImageNet[36] weights with fine tuning. The Adam optimizer[37] with a learning rate of 1e–7 was employed. Data augmentation was performed, comprising transformations such as scaling, translations, flipping, translations, contrast normalization, rotations, and brightness alterations. The loss function used for training is binary cross-entropy. Training was stopped when no increase in the area under the receiver operating characteristic curve (AUROC) computed at the frame-level on the tuning set was observed after 10 consecutive epochs.

Appendix 3

Model evaluation

A 5-fold cross-validation scheme (Figure 10) was used to train and evaluate the QC and UCS models. The splitting of the data set is performed as follows:

1. We split the full data set in 5 randomly selected subsets by applying a site-level constraint. This constraint ensures that data from a single site does not feature in multiple subsets.
2. We divided the 5 random subsets into three groups: training (3 subsets), tuning (1 subset), and testing (1 subset). The model is trained using the training group, the tuning subset is used to select an optimal set of hyperparameters and the final evaluation is performed on the testing set.
3. We repeat the step 2 keeping the testing group fixed, and using a different validation group.
4. We repeat the whole procedure 5 times. Every time a different proportion of the data set is used as either the development or testing data set.

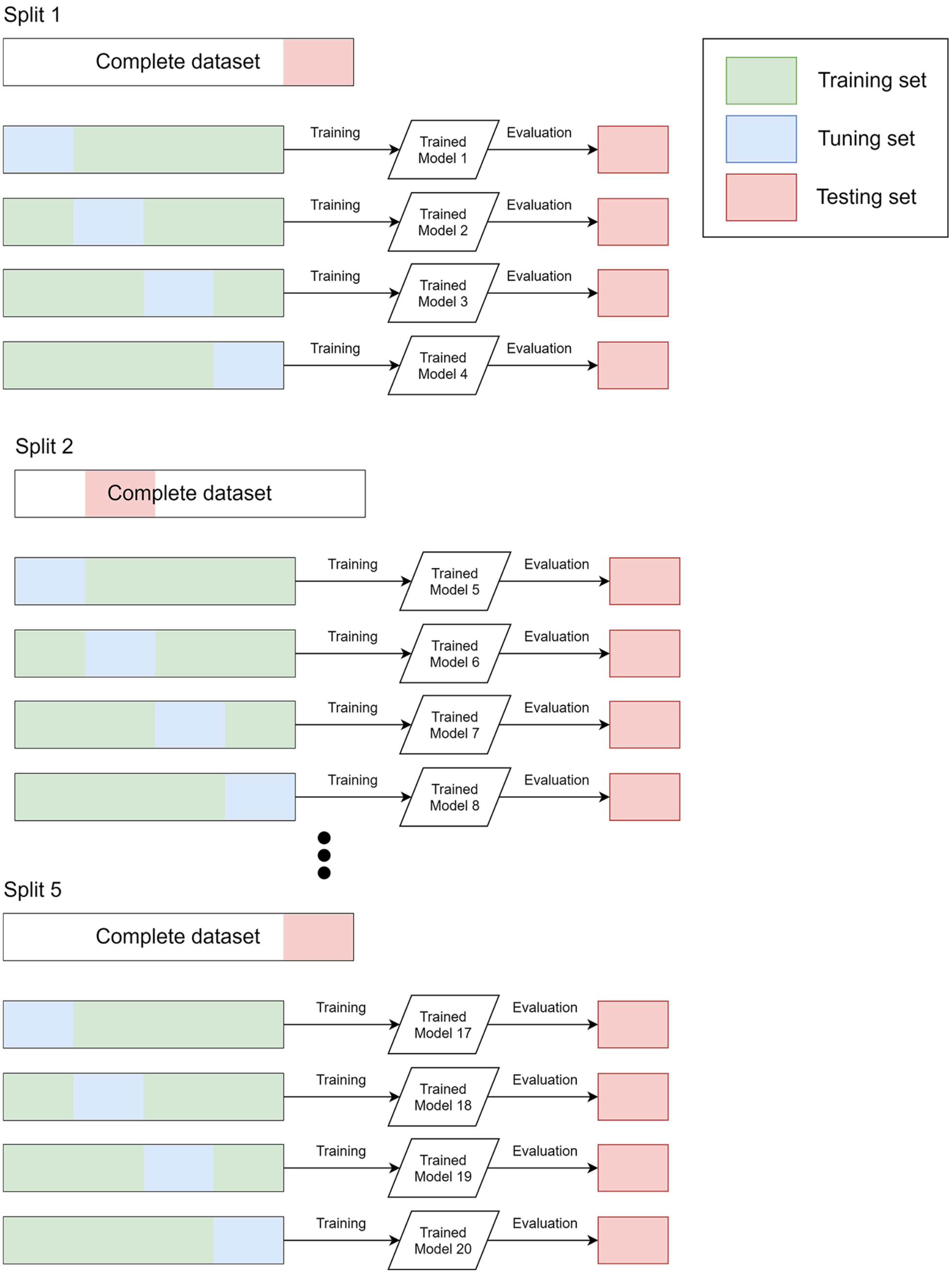


Figure 10. Schematic representation of the splitting strategy used to train and evaluate all the models presented in this work. Five different splits of the full data set are created. For each split we conduct 4 experiments, each one using a different part of the data set as the tuning data set. For each experiment, we train and evaluate 20 models in total, each one with its own training and validation set. For each experiment, the testing set is used only for the final evaluation and not for the selection of the optimal model or hyperparameter tuning.

Following the procedure described above, we perform 20 independent runs for each one of our experiments, and we report mean and standard deviation values. We assessed the performance of the models by computing the AUROC over the testing sets. The QC model was trained and evaluated using only data from the Eucalyptus clinical trial, whereas the end-to-end UC prediction model was trained on data from Hickory and Laurel. AUROC scores were computed using Python and the scikit-learn library.