Fully automated endoscopic disease activity assessment in ulcerative colitis

Heming Yao, BS,[1] Kayvan Najarian, PhD,[1,2,3,4,5] Jonathan Gryak, PhD,[1,5] Shrinivas Bishu, MD,[7] Michael D. Rice, MD,[7] Akbar K. Waljee, MD, MSc,[6,7,8,9] H. Jeffrey Wilkins, MD,[10] Ryan W. Stidham, MD, MS[1,6,7]

Ann Arbor, Michigan; Plymouth Meeting, Pennsylvania, USA

Background and Aims: Endoscopy is essential for disease assessment in ulcerative colitis (UC), but subjectivity threatens accuracy and precision. We aimed to pilot a fully automated video analysis system for grading endoscopic disease in UC.

Methods: A developmental set of high-resolution UC endoscopic videos were assigned Mayo endoscopic scores (MESs) provided by 2 experienced reviewers. Video still-image stacks were annotated for image quality (informativeness) and MES. Models to predict still-image informativeness and disease severity were trained using convolutional neural networks. A template-matching grid search was used to estimate whole-video MESs provided by human reviewers using predicted still-image MES proportions. The automated whole-video MES workflow was tested using unaltered endoscopic videos from a multicenter UC clinical trial.

Results: The developmental high-resolution and testing multicenter clinical trial sets contained 51 and 264 videos, respectively. The still-image informative classifier had excellent performance with a sensitivity of 0.902 and specificity of 0.870. In high-resolution videos, fully automated methods correctly predicted MESs in 78% (41 of 50, $\kappa = 0.84$) of videos. In external clinical trial videos, reviewers agreed on MESs in 82.8% (140 of 169) of videos ($\kappa = 0.78$). Automated and central reviewer scoring agreement occurred in 57.1% of videos ($\kappa = 0.59$), but improved to 69.5% (107 of 169) when accounting for reviewer disagreement. Automated MES grading of clinical trial videos (often low resolution) correctly distinguished remission (MES 0,1) versus active disease (MES 2,3) in 83.7% (221 of 264) of videos.

Conclusions: These early results support the potential for artificial intelligence to provide endoscopic disease grading in UC that approximates the scoring of experienced reviewers.

Abbreviations: AI, artificial intelligence; AUC, area under the curve; CI, confidence interval; CNN, convolutional neural network; FPS, frame per second; IBD, inflammatory bowel disease; MES, Mayo endoscopic score; UC, ulcerative colitis.

DISCLOSURE: *A provisional patent has been filed on behalf of Drs Stidham, Najarian, Yao, and Gryak on the methods presented in this study, with technology elements licensed to AMI, Inc. Dr Wilkins is an executive of Lycera Corp and holds stock options. Dr Stidham has been a consultant for Abbvie, Janssen, Merck, Takeda, and Corrona.*

0016-5107/$36.00
https://doi.org/10.1016/j.gie.2020.08.011

Received April 2, 2020. Accepted August 11, 2020.

Current affiliations: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan (1); Department of Emergency Medicine, University of Michigan, Ann Arbor, Michigan (2); Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan (3); Michigan Center for Integrative Research in Critical Care (MCIRCC), University of Michigan, Ann Arbor, Michigan (4); Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, Michigan (5); Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), University of Michigan, Ann Arbor, Michigan (6); Division of Gastroenterology and Hepatology, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan (7); Division of Gastroenterology and Hepatology, Department of Internal Medicine, Veteran Affairs Ann Arbor Health Care System, Ann Arbor, Michigan (8); Veterans Affairs Center for Clinical Management Research, Ann Arbor, Michigan (9); Lycera Corporation, Plymouth Meeting, Pennsylvania, USA (10).

Reprint requests: Ryan W. Stidham, MD, MS, University of Michigan School of Medicine, 1500 East Medical Center Drive, 3912 Taubman Center, Ann Arbor, MI 48109.

If you would like to chat with an author of this article, you may contact Dr Stidham at ryanstid@med.umich.edu.

INTRODUCTION

Endoscopic measurement of mucosal injury is an important component of disease severity assessment in ulcerative colitis (UC). Although existing and emerging biomarkers, such as fecal calprotectin and histopathologic scoring, provide additional measures of biological disease activity, endoscopy continues to serve as the reference for objective disease assessment. The importance of endoscopy cannot be understated; it is a principal component of definitions for disease severity and therapeutic response used in both the assessment of investigational medications and the day-to-day decision making for the patient with UC. As a result, routine endoscopy to assess disease status is recommended in the recently published American College of Gastroenterology clinical management guidelines, the STRIDE (Selecting Therapeutic Targets in Inflammatory Bowel Disease) international consensus statement, and by regulators in the setting of clinical trials.[1-3]

Efforts to operationalize grading of UC severity have resulted in multiple scoring systems; the Mayo endoscopic score (MES) is the most commonly used, likely owing to its simplicity and physician familiarity.[4] The MES, developed in the 1980s, is a 4-level scale of severity (scored 0-3) with higher scores reflecting increasing disease severity based on features including erythema, erosions, ulcerations, and bleeding.[5] Beyond assessing therapeutic effect, low or reduced MESs are associated with a lower risk of future colectomy and clinical relapse.[6,7] However, the subjectivity of qualitative image interpretation introduces problems of inter- and intraobserver variability, as well as treatment and selection bias, threatening the accuracy and reproducibility of these important assessments. When asked to grade overall disease severity using endoscopic videos, 10 specialists in inflammatory bowel disease (IBD) had 78% agreement when severe disease was present, but only 37% and 27% agreement for moderate disease and normal, respectively.[8]

To combat these problems, central reading of endoscopy by experienced clinicians who are trained to use endoscopic grading instruments has been adopted in nearly all therapeutic clinical trials in UC and was first used in the mid-2000s.[9] Central reading has important clinical relevance as demonstrated in clinical trials where eligibility of patients with UC, determination of therapeutic response, and ultimately the perceived success or failure of an investigational agent differed based on local expert versus blinded central reader endoscopic assessment.[10] Yet, central review can still be challenged by several issues, including the subtle distinctions between disease severity grades, the short supply of trained reviewers, and the time needed for high-quality adjudicated video review.

Advances in machine learning methods, colloquially referred to as artificial intelligence (AI), provide a means to address the inherent subjectivity in human image interpretation. Gastroenterology has seen a rapid emergence of AI methods designed to replicate expert endoscopic interpretation, principally in the colonic adenoma and polyp recognition space.[11,12] In previous work, we have shown that deep learning techniques can classify the MES in still endoscopic images with a similar performance to experienced human reviewers.[13] However, generating a fully automated summary of MES disease severity using an entire full-motion colonoscopy video is hindered by several barriers. Disambiguation of UC disease activity from debris, stool, poor image quality, and interventional versus endogenous tissue damage is a task readily performed by experienced gastroenterologists but poses a formidable interpretation problem for computational systems. Here, we investigated methods for analysis of unaltered full-motion videos, aided by methods to distinguish informative versus noninformative portions of a video, to automatically generate MESs for patients with UC.

METHODS

Study cohorts

Internal high-resolution cohort. Ethical review of the study protocol was performed by the local Institution Review Board. Consecutive patients presenting for clinically indicated colonoscopy to evaluate established UC were recruited for participation in the local development cohort. UC diagnosis was defined using the following factors: 2 administrative diagnosis codes for UC (ICD-9 or ICD-10) on 2 separate encounters, previous histologic UC diagnosis, and the use of at least one UC medication.[14] Patients with a diagnosis of Crohn's disease, indeterminate colitis, ileoanal pouch anastomosis, colostomy, ileostomy, or other bowel resection, or known infectious colitis or dysplasia were excluded from study participation. Each colonoscopy collected was from a unique patient. In participating patients, colonoscopy videos were recorded at 1920 × 1080 high resolution, 10-bit color depth, and 60 frames per second. All colonoscopies were performed using a CF-HQ190 or PCF-H190 colonoscope and CLV-190 image processors (Olympus Corporation, Inc, Tokyo, Japan). The MES was obtained from local endoscopy videos by 2 local central reviewers blinded to clinical status.

External multicenter clinical trial testing cohort. The automated endoscopic scoring workflow developed using internal high-resolution videos was tested on external videos from the LYC-30937-EC study, an international phase 2 randomized clinical trial of an investigational oral therapy for moderate to severe UC (clinical trial registration number: NCT02762500). In this study, both full colonoscopy and sigmoidoscopy were recorded at week 0 and 8 using local equipment and recording methods. The variation in endoscopic site, equipment, and recording techniques provide an advantage for testing the

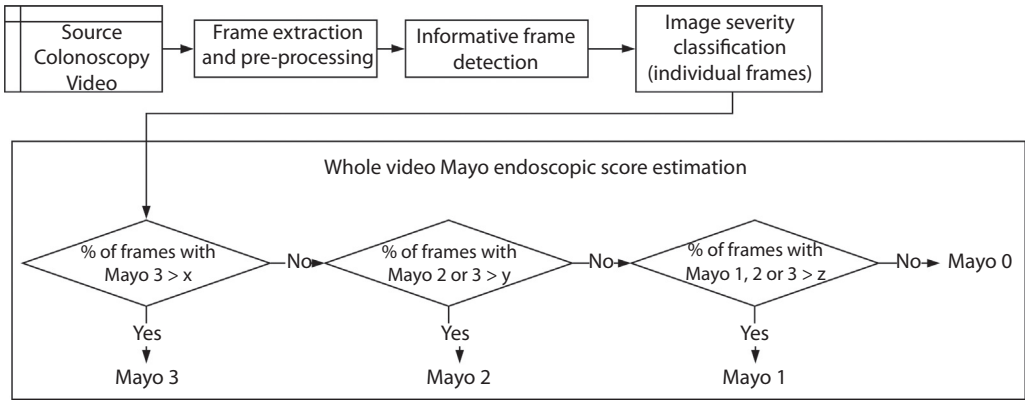


Figure 1. Schematic of a pilot fully automated Mayo endoscopic grading process. Unaltered endoscopic source videos flow through several image processing modules to generate a whole video summary Mayo endoscopic score prediction. Videos are converted into a still frame stack and are cropped and scaled for image uniformity. After preprocessing, images are passed to a classifier to separate informative (gradable) versus noninformative (ungradable) images. Informative images are then graded for disease severity using the pre-trained Mayo endoscopic subscore classifier. Finally, the relative proportion of predicted disease severity grades for all informative images within the video are used to estimate the whole video Mayo endoscopic score.

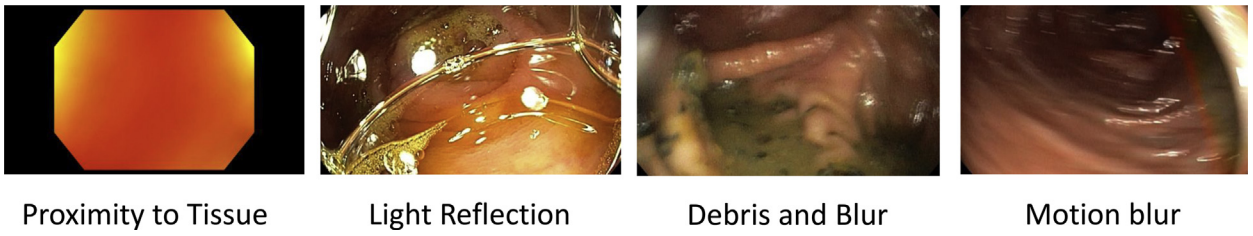


Figure 2. Detection of informative versus noninformative images in endoscopic videos to identify portions of videos suitable for disease severity assessments. Real-world unaltered endoscopic video is composed of a large fraction of images that are unsuitable for disease severity grading. Common confounders and noise that can influence automated disease severity grading include those shown in the figure, as well as over- and underexposure of light sources. The performance of automated informative image detection systems is detailed in Table 2.

performance of automated analysis methods in real-world settings. Endoscopic videos were centrally reviewed for MES by external reviewers as part of the original study protocol and served as the ground truth. The investigators in the presented analysis did not participate in the central review scoring process of external videos from the clinical trial. Clinical trial videos were not used in automated modeling or workflow development.

Fully automated endoscopic analysis workflow. A schematic for automated endoscopic video analysis architecture, comprising sequential analysis modules for informative image detection, still-image severity classification, and then overall MES estimation, is shown in Figure 1. Endoscopic videos were first segmented into 1 frame per second (FPS) still-image stacks. Images underwent preprocessing and were automatically cropped and padded into a square shape for image consistency without skewing image proportions, as well downscaling to 256 × 256 resolution for inputs into convolutional neural networks (CNNs). Images also underwent random transformations of rotation, zoom, sheer, and vertical and horizontal orientation to improve the variability of the dataset and prevent overfitting. Human labeling of images was performed on the source still images, not the transformed images. The intention of the overall workflow was fully automated MES generation for an unseen video. Training and testing of individual analysis modules were performed using the local high-resolution video dataset with validation of fully automated workflows using the external clinical trial dataset.

Informative image classifier development. To remove images unsuitable for UC disease activity assessment, a deep learning CNN approach was used to automatically distinguish informative (gradable) from noninformative (nongradable) images. Noninformative images are defined here by several qualitative characteristics, including (1) the camera being too close to the mucosa, (2) obscured mucosa due to insufficient bowel preparation, (3) excessive motion blur, (4) over- or under-lighting exposure (Fig. 2). Development of the informative frame detector model used 30 videos from the internal high-resolution cohort at 1 FPS, qualitatively classifying each image as informative (gradable) or noninformative (nongradable).

In this study, Python 3.7 with TensorFlow libraries was used for model generation. The informative image CNN model development used the Inception-V3 architecture initialized by pre-trained weights using ImageNet.[15] The Inception V3 architecture, which is a 42-layer CNN, has

been successfully applied to extensive image recognition tasks.[16] Adaptive moment estimation with a learning rate of 10^{-6} was used to fine-tune the network.[17] In model training, optimization, and testing, a 5-fold cross-validation was used to remove the bias from data splitting, because images from the same video can only be present in 1 fold. The hyper-parameters we used in this study were from our previous work.[18] To evaluate the informative classifier performance on real-world videos of variable video quality, 10 colonoscopy videos from the external clinical trial video dataset were randomly selected and annotated for image quality as an independent test dataset.

Still endoscopic image disease severity classification. Informative frames were passed to the UC still-image severity classifier to estimate the MES for each informative image. A pre-existing MES classifier was used, as detailed in our previous work, which has performance characteristics that approximate the still-image MES scoring agreement of paired experienced reviewers.[13] The MES still-image classifier was modeled using a dataset including approximately 3000 UC patients and 16,000 still images that were labeled by 2 IBD specialists with clinical trial experience; score disagreements were adjudicated by a third reviewer. The classifier was designed to separate Mayo 0, 1, 2, 3 endoscopic disease severity levels using a still image. This classifier was designed using the same CNN development methods as described in the previous section.

Full video MES estimation. For each video, the percentages of informative frames classified as Mayo 0, 1, 2, or 3 were calculated. The overall summary Mayo score was inferred based on the proportion of frames in a video for each given MES class (eg, Mayo 3 comprises 12% of frames, Mayo 2 comprises 25% of frames, Mayo 1 comprises 23% of frames, and Mayo 0 comprises 40% of frames). The highest Mayo score meeting the threshold proportion of frames in a video was selected as the overall Mayo score (Fig. 1). The MES proportion thresholds for the overall summary score were determined using a template-matching grid search where the threshold proportions of MESs in a video were matched to the overall Mayo score provided by expert review of the entire video. The rationale for requiring a threshold number of video frames to validate the presence of a severity class is to address potential misclassifications in single-frame severity grading or confounding from other causes that could have an impact on the overall scoring. It also corresponds to the fact that a human reviewer does not consider single frames in isolation but actually many seconds worth of video to determine the severity present. Similar to the informative frame detection, 5-fold cross-validation was used to evaluate the proposed summary Mayo score estimation method on the internal cohort. In each round, 4 rounds were used to train the model, and the remaining fold was used as the unseen test data.

Statistical analysis and testing on the external clinical trial video dataset. The performance of image classification models using the internal high-resolution videos were reported as the accuracy, sensitivity, specificity, precision, F1 score, and AUC on the 5-fold cross-validation testing fold. Both the average and standard deviation of these evaluation metrics from 5 folds were used to eliminate bias from data split and measure the robustness of the model. Confusion matrices for the whole-video endoscopic score and the predicted whole-video score were generated to compare paired human reviewers and reviewer to automated scoring. Agreement between human reviewers and predicted scores used Cohen's kappa coefficient with quadratic weighting, which has similar performance to the intraclass correlation coefficient.[19]

After classifier development and optimization using the internal video dataset, the MES workflow pipeline was tested on the external clinical trial dataset videos to assess the generalizability of the proposed system. The external videos were unaltered from the source clinical trial video file and were not used in model development.

RESULTS

Study population characteristics for local high-resolution and external testing video sets

The local high-resolution video set contained 51 videos, whereas the testing video set contained 264 videos from 157 candidate patients (Table 1); note that only 124 patients ultimately met the clinical trial enrollment criteria but screen failure videos were included to increase the proportion of low disease activity in the dataset. The clinical trial videos were collected from 72 sites (United States, Canada, and 5 European countries); 77.4% of videos were collected in Europe. The local video set had a more even distribution of endoscopic severity (MES 0, 1, 58.8%; MES 2, 3, 41.2%) compared with the external clinical trial testing video set (MES 0, 1, 16.3%; MES 2, 3, 83.7%; $P < .0001$). In addition, the clinical disease severity as assessed by the total Mayo score and corticosteroid use was more severe in the clinical trial test set. These differences are unsurprising because clinical trial subject recruitment skews toward more severe disease activity.

Informative versus noninformative image classifier performance. In the internal high-resolution video set, a total of 34,810 frames were extracted and classified as "informative" versus "noninformative." The CNN model demonstrated excellent performance for distinguishing informative versus noninformative images based on the summary results from the 5-fold cross-validation (Table 2). The informative image classifier performance had an average AUC of 0.961 (0.010) over the 5 folds with an accuracy, sensitivity, and specificity of 0.876 (0.010), 0.902 (0.036), and 0.870 (0.030), respectively, for

TABLE 1. Patient characteristics in local developmental and external clinical trial video sets

Characteristic	Local video developmental training set (n = 51)	Clinical trial test video set (n = 124)	*P* value
Age (years), mean (SD)	43.5 (15.4)	41.5 (12.8)	.378
Sex, n (%) female	22 (43.1)	52 (41.9)	.884
BMI (kg/m^2), mean (SD)	27.1 (5.5)	25.7 (4.7)	.091
Disease duration (years), mean (SD)	8.4 (7.4)	7.7 (6.8)	.547
Total Mayo score, mean (SD)	3.9 (2.7)	7.9 (1.6)	<.001
C-reactive protein ≥5 mg/L, n (%)	n/a*	61 (49.6)	
Fecal calprotectin range, n (%)			
≤250 ug/g	n/a*	27 (22.1)	
>250 to ≤500 ug/g	n/a*	19 (15.6)	
>500 ug/g	n/a*	76 (62.3)	
Medication use, n (%)			
None	0 (0.0)	2 (1.6)	.854
5-ASA	34 (66.7)	109 (87.9)	<.001
Corticosteroids	8 (15.7)	68 (54.8)	<.001
Thiopurines	15 (29.4)	32 (25.8)	.625
Biologic exposure	18 (35.3)	25 (20.2)	.035
Race, n (%)			
American Indian or Alaskan Native	0 (0.0)	0 (0.0)	.999
Asian	1 (2.0)	1 (0.8)	.514
Black or African American	4 (7.8)	3 (2.4)	.096
Native Hawaiian or Pacific Islander	0 (0.0)	0 (0.0)	.999
White	46 (90.2)	119 (96.0)	.135
Other	0 (0.0)	1 (0.8)	.788
Ethnicity, n (%) Hispanic or Latino	1 (2.0)	7 (5.6)	.289

SD, Standard deviation; *BMI*, body mass index; *ASA*, aminosalicylic acid.
*Prospective C-reactive protein and fecal calprotectin levels were inconsistently available in the developmental video set.

separating gradable from ungradable images using the gastroenterologist image quality labels as the ground truth. Learning curves detailing classification performance and efficacy based on the number of images used for informative image and disease severity classifier modeling are shown (Supplementary Fig. 1, available online at www.giejournal.org). From the learning curve, the numbers of training samples in our study are sufficient and adding more training sample may result in only limited improvement. We tested the informative image classifier on images from colonoscopy videos in the external clinical trial video dataset, because the external dataset was not used in training (Table 2). This dataset is more challenging compared with the high-resolution dataset because the videos are from different sources with varied recording conditions. The classifier achieved an average AUC of 0.930, with an accuracy, sensitivity, and specificity of 0.844, 0.834, 0.851, providing similar performance compared with testing on high-resolution videos.

Across all 51 high-resolution videos, the median portion of informative frames per endoscopy was 59.3% (standard deviation, 14.4%) with a maximum of 82.5% and minimum of 28.8%. In the external clinical trial dataset, containing standard or resolution videos with variable recording equipment, the median portion of informative video was 43.1% (standard deviation, 17.5%) with a range between 85.5% and as low as 3.7%.

Whole-video fully automated endoscopic scoring using local high-resolution videos. The automated MES system exhibited very good agreement with gastroenterologist scoring (κ = 0.84; 95% confidence interval [CI], 0.75-0.92) and correctly predicted the exact MES in 40 of 51 (78%) high-resolution videos (Table 3). MES severity grading thresholds for Mayo 1, 2, and 3 of 7%, 6%, and 6%, respectively, were used for entire-video score prediction. Unsurprisingly, disagreement was concentrated in mild disease severity classes, including Mayo 1, where 5 of 9 cases were classified as Mayo 2 and 1 of 9 was classified as Mayo 0. Paired gastroenterologist reviewers agreed on exact MES in 84.3% of cases (κ = 0.95), similarly with disagreement concentrated in the intermediate Mayo 1 and 2 classes (Supplementary Table 1, available online at

TABLE 2. Informative versus noninformative image classification performance in full-motion endoscopic video of ulcerative colitis

	Accuracy	Sensitivity	Specificity	Area under the curve	F1 score	Precision	Average precision
Classifier performance using a uniform high-resolution video test set							
Fold 0	0.88	0.85	0.908	0.962	0.84	0.853	0.941
Fold 1	0.885	0.907	0.856	0.97	0.775	0.684	0.948
Fold 2	0.862	0.944	0.831	0.945	0.837	0.754	0.88
Fold 3	0.868	0.888	0.889	0.961	0.86	0.849	0.944
Fold 4	0.883	0.923	0.868	0.967	0.859	0.812	0.948
Average	0.876	0.902	0.87	0.961	0.834	0.79	0.932
(std)	(0.010)	(0.036)	(0.030)	(0.010)	(0.035)	(0.064)	(0.026)
Classifier performance using an external variable video quality test set							
External video set	0.844	0.834	0.851	0.93	0.804	0.831	0.91

Classification performance for automated determination of informative versus noninformative images within an endoscopy video, based on human reviewer assessment of the ability to grade the UC disease severity of a still image. The classifier was trained on uniform high-resolution video; the high-resolution test set performance was consistent across the 5 analysis folds for the 34,810 reviewed video frames. Testing the informative classifier on 10 randomly selected external videos of variable quality, which were not included in training, yielded similar performance.

TABLE 3. Automated Mayo endoscopic scoring performance using the local developmental high-resolution video set with and without accounting for informative versus noninformative images

	Automated predicted score			
	Mayo 0	Mayo 1	Mayo 2	Mayo 3
Mayo endoscopic scoring using informative image classifiers				
Reference scoring				
Mayo 0	19	2	0	0
Mayo 1	1	3	5	0
Mayo 2	0	0	7	0
Mayo 3	0	0	3	11
Mayo endoscopic scoring without informative image classifiers				
Reference scoring				
Mayo 0	17	1	2	1
Mayo 1	1	1	7	0
Mayo 2	2	0	5	0
Mayo 3	1	0	3	10

Incorporating a method to automate informative image classification improved the performance of automated endoscopic scoring in the developmental video set using high-resolution videos. Without the informative versus noninformative classifier, the correctness of automated Mayo endoscopic score was 64.7% with a modest agreement between human reviewer and automated methods for exact score ($\kappa = 0.63$; 95% confidence interval, 0.52-0.89). The use of the informative image classifier improved overall fully automated video score correctness to 78.4% with very good agreement on exact score ($\kappa = 0.84$; 95% confidence interval, 0.75-0.93). Paired reviewer agreement on exact Mayo endoscopic score was numerically better at 84.3% ($\kappa = 0.95$).

www.giejournal.org). Highlighting the performance improvements gained by preventing images inappropriate for disease grading from being included in the overall severity estimation, removing the noninformative classifier from the MES prediction process resulted in worse accuracy and agreement compared with experienced gastroenterologist reviewers (correctness of 64.7%; agreement $\kappa = 0.63$; 95% CI, 0.52-0.89).

Automated endoscopic scoring using unaltered multicenter clinical trial videos. Using the same informative image and disease severity classification processes, we analyzed 264 videos, including both screening and follow-up videos from the LYC-30937-EC study. Agreement between the automated predicted MESs and reference MESs provided by external central review was moderate on unadjusted analysis ($\kappa = 0.59$; 95% CI, 0.46-0.71); 57.1% (151 of 264) of videos were correctly graded based on the central review score provided. Automated endoscopic analysis was within 1 MES severity level of the score provided by central reviewers in 93.5% (247 of 264) of videos. Fully automated methods correctly separated Mayo 0, 1 versus Mayo 2, 3 endoscopic severity in 83.7% (221 of 264) videos compared with the reference central reviewer score. The comparative performance of fully automated whole-video MES performance between local high-resolution videos and external multisite videos are shown

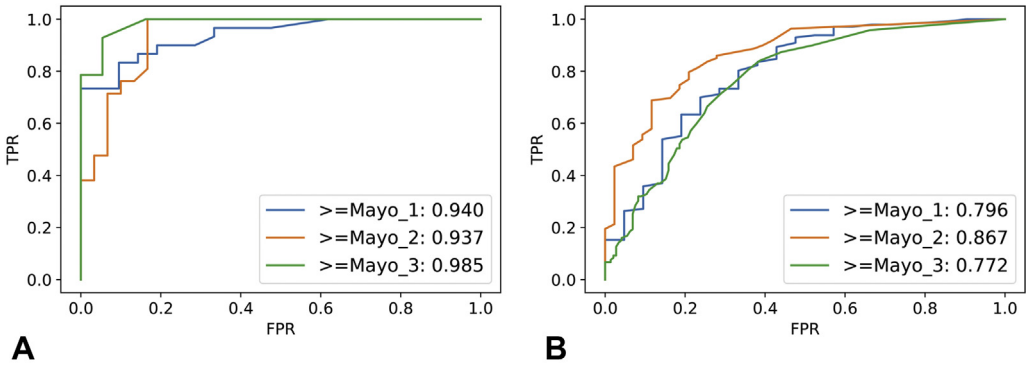


Figure 3. Relative performance of fully automated whole-video Mayo endoscopic score using high-resolution and variable-quality external videos. Comparatively, overall score prediction performance was superior using high-resolution videos (**A**) compared with a mixture of variable-quality external endoscopic videos sourced from an international multisite clinical trial (**B**). Ordinal characteristics are shown for the automated process predicting progressive increases in disease severity. Video variability features include image resolution, fragmentation of videos, and inconsistent frequency and count of mucosal biopsies. Exact score performance confusion matrices for high-resolution and variable-quality clinical trial videos are shown in Tables 3 and 4, respectively. *FPR*, False positive rate; *TPR*, true positive rate.

in Figure 3. We explored the impact of video sampling rate and found that compared with 30 FPS, lower frame rates of 15, 5, and 1 FPS had a marginal reduction in predicted MES accuracy of 56.8%, 55.7%, and 54.9%, respectively.

The accuracy, sensitivity, and specificity characteristics varied by each MES level; these are listed in Table 4. Qualitative misclassification analysis of the 17 of 264 (6.4%) automated predicted MESs that were 2 levels different than central review scores was performed. Overestimated disease severity (eg, Mayo 0 predicted as 2 or 3) included extensive biopsy sampling with resulting mucosal bleeding, which was interpreted as severe disease. Underestimated scores (eg, Mayo 2 predicted as 0) had short segments of severe disease qualifying the patient as a high endoscopic severity grade, although the severe disease was a small fraction of the disease burden.

Considering the potential ambiguity between neighboring MES grades, even among experienced clinicians and reviewers, we explored the performance of automated endoscopic grading when adjusting for scenarios where reviewers disagreed on MES grade. The purpose of this exploratory adjusted analysis was to avoid penalizing automated systems for the subjectivity of "correctness" in ground truth scoring. Only 169 of 264 (64.0%) videos underwent dual central reader review based on the clinical trial video review protocol. Where dual video review was performed, expert reviewers agreed on the exact score in 82.8% (140 of 169) of videos ($\kappa = 0.78$; 95% CI, 0.71-0.86), with disagreement by 1 level (eg, Mayo 0 vs Mayo 1) in 14.2% (24 of 169) of videos. When defining an automated MES prediction matching the MES provided by either expert reviewer as "correct" (eg, reviewer A scoring Mayo 2 and reviewer 2 scoring Mayo 3), the automated MES was correct in 69.5% (107 of 169) of cases, compared with 57.1% correct without adjusting for central reviewer disagreement. Interestingly, disagreement between reviewers was not associated with increased odds of an incorrect automated score prediction (odds ratio, 1.60; 95% CI, 0.70-3.12), suggesting factors beyond human reviewer disagreement also contributed to automated MES score prediction failures.

DISCUSSION

Classifying the presence or absence of a disease, finding, or severity using static images is becoming increasingly available in gastroenterology. However, automating the cumulative assessment of an entire endoscopic video from an experienced specialist presents additional challenges. Here, we show the progress toward full automation of endoscopic grading in UC using deep learning techniques. The addition of a new layer of information, namely which portions of video are informative (gradable) versus noninformative (ungradable) improved the performance of automated endoscopic scoring. Testing automated MES scoring on externally sourced videos demonstrated encouraging results for the potential of computational endoscopic analysis. Although exact disease severity grading based on current MES conventions requires more work, good performance of CNN-based separation of Mayo 0-1 versus 2-3 classes, a major boundary for endoscopic response in UC, offers near-term applications in assisting clinical trial enrollment and new concepts in value-based care. However, the performance is imperfect and new methods beyond image classification alone will be needed to manage confounders that affect image analysis performance, including biopsy-related bleeding, variations in video recording quality, and addressing image compression artifacts.

Increasingly, neural network classification methodologies are being applied to categorize endoscopic, histologic, and radiologic images into known groupings (eg, diseased

TABLE 4. Performance of fully automated Mayo endoscopic scoring using unaltered videos from a multicenter clinical trial

	Mayo 0	Mayo 1	Mayo 2	Mayo 3
Automated Mayo endoscopic scoring agreement with central expert review scores				
Reference scoring				
Mayo 0	9	8	2	2
Mayo 1	4	10	6	2
Mayo 2	2	18	48	34
Mayo 3	1	8	27	83
Automated Mayo endoscopic scoring performance using central review scores as ground truth				
Accuracy	0.947	0.888	0.678	0.711
Sensitivity	0.500	0.800	0.538	0.667
Specificity	0.972	0.894	0.782	0.747

Automated methods predicted the same MES as external central reviewers in 57.1% (151/264) of videos ($\kappa = 0.59$). When dual central review was performed, reviewers agreed on exact MES in 82.8% (140/169) of videos ($\kappa = 0.78$). When accounting for central reviewer disagreement, automated MES prediction was correct in 69.5% (107/169) of videos.
MES, Mayo endoscopic score.

vs normal). On the path of progress toward reliable fully automated endoscopic assessments in both IBD and other GI diseases, it is becoming apparent that image classification alone (eg, cat vs dog, adenoma vs hyperplastic polyp, mild vs severe mucosal inflammation) will only take the field so far in the quest to approximate and improve upon expert inference and medical judgment. As is the case in the presented work focused on UC, contextual awareness is of critical importance when determining the relationships, meaning, and importance between sequenced images relative to an overall disease assessment. Despite being a simple and intuitive principle for experts, the context of useful and nonuseful visual information involves complex concepts to program into machine understanding. Considering that a 20-minute video contains 36,000 individual frames (1200 frames at 1 FPS) and approximately 60% of frames may be ungradable for disease severity, automating the judgment of informative versus noninformative images is essential for a practical scoring workflow. As shown, informative versus noninformative image awareness substantially improves the performance of endoscopic analysis systems with the flexibility to detect images unsuitable for grading even in unseen videos with low quality.

On first appraisal, these early efforts toward fully automating UC severity grading highlight that much more work is needed to replicate the intricacies of expert endoscopic judgment. However, recognizing the difficulty of analyzing the videos used to test these methods demonstrates progress in image analytics. The clinical trial videos had substantial variability in terms of (1) video quality resolution, compression, and color gamut, (2) both sigmoidoscopies and colonoscopies were used, (3) the frequency of mucosal biopsies, and (4) the variable duration and tempo of the endoscopies. The external video variability should be considered a strength of the study, setting a high-performance expectation for a system that must handle a myriad of confounding common real-world variables. Although the automated MES system has good performance using high-quality endoscopic videos, systems of value will need to prove good performance in all clinical scenarios and environments. In addition, noting the variation in endoscopic video quality, these results highlight the importance of standardization of video acquisition and digitization practices in clinical trials and clinical care.[20]

This work was subject to several limitations. The ground truth for endoscopic disease severity assessments is inherently subjective. Arguments could be made for expanding the number of expert labelers for still images and endoscopic videos. However, there is no perfect reference for endoscopic scoring, and central reading does not completely eliminate bias, disagreement, or variability of ground truth disease severity grading. Similarly, despite the source MES still-image disease severity definitions being the product of an adjudicated dataset, bias on score opinion is possible. We believe that using an external video set for testing, unseen in any part of development, mitigated the risk of these possible biases, offering a challenging test of automated endoscopic scoring performance. We also acknowledge our best "ground truth" for endoscopic assessment used to train and judge automated systems has substantial remaining limitations. Efforts to date to improve standardization and reproducibility of endoscopic assessment have included (1) establishing descriptions of semiquantitative severity score criteria, (2) the use of central readers detached from management decisions, and (3) training methods for reviewers to improve uniformity. Increasingly, our ground truth for endoscopic feature evaluation and scoring may need to be reconsidered given increasing availability of computational methods for more discrete and reproducible image assessment.

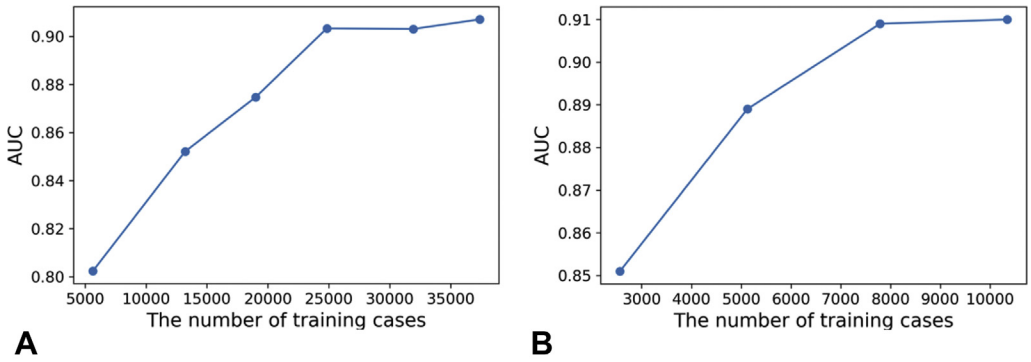
Another important limitation was the difference in disease severity between the patients in the developmental and clinical trial videos, who expectantly contained a

higher proportion of moderate to severe disease. We believe the disease severity distribution typical of clinical trials is justified by the expectation that these populations are likely where video analysis will be first applied. Future development and validation methods will benefit from evenly distributed disease severity datasets to be of the most value in research and clinical care. Finally, intraobserver variation for automated methods could not be addressed in this study. Although automated methods demonstrated a perfect reproduction of MES when processing the same video file, the appropriate assessment of computed MES variation would require repeat colonoscopy on the same patient.

In conclusion, although existing methods are premature for immediate deployment, these early results support the potential for AI to provide endoscopic disease severity grading in IBD. Neural network models for disease activity scoring would provide broad accessibility to unbiased and reproducible disease assessments, because they can run efficiently on a consumer desktop computer with an upgraded graphics card (eg, NVidia Tesla V100 processing speed is 5 minutes per video). Compared with central review costing hundreds to thousands of dollars for each endoscopic video, automated computational scoring approaches are likely to provide a more cost-effective means for therapeutic trials, research, and clinical practice. This work highlights key barriers to overcome to improve endoscopic analysis performance, including disambiguation of endogenous versus interventional tissue injury and developing an improved awareness of disease distribution that is part of ongoing work. Emerging instruments such as the UC endoscopic index of activity (UCEIS) may offer advantages over MES.[8] Although more complex than the MES, UCEIS grading of individual visual components of UC activity (eg, vascularity, ulceration, bleeding) could reduce the ambiguity of the MES that challenge automated computational severity assessments.[21] In addition, as histology becomes increasing relevant in UC severity assessment, deep learning methods are being explored to infer histologic score based on endoscopic appearance alone.[22] Finally, the full potential of AI methods will not be realized though replicating, but instead reimagining disease assessment by moving beyond arbitrary classifications of severity. Machine learning methods in gastroenterology are in their infancy but are maturing rapidly. AI has begun to demonstrate expert level judgment using cleaned and curated data and images, and is now beginning to show promise for understanding endoscopic video.

REFERENCES

1. Rubin DT, Ananthakrishnan AN, Siegel CA, et al. ACG clinical guideline: ulcerative colitis in adults. Am J Gastroenterol 2019;114:384-413.
2. Peyrin-Biroulet L, Sandborn W, Sands BE, et al. Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE): determining therapeutic goals for treat-to-target. Am J Gastroenterol 2015;110:1324-38.
3. US Food and Drug Administration. Ulcerative colitis: clinical trial endpoints guidance for industry. Available at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM515143.pdf. Accessed June 18, 2019.
4. Mohammed Vashist N, Samaan M, Mosli MH, et al. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Syst Rev 2018;1:CD011450.
5. 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.
6. Colombel JF, Rutgeerts P, Reinisch W, et al. Early mucosal healing with infliximab is associated with improved long-term clinical outcomes in ulcerative colitis. Gastroenterology 2011;141:1194-201.
7. Barreiro-de Acosta M, Vallejo N, la Iglesia de D, et al. Evaluation of the risk of relapse in ulcerative colitis according to the degree of mucosal healing (Mayo 0 vs 1): a longitudinal cohort study. J Crohns Colitis 2016;10:13-9.
8. Travis SPL, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut 2012;61:535-42.
9. Sandborn WJ, Regula J, Feagan BG, et al. Delayed-release oral mesalamine 4.8 g/day (800-mg tablet) is effective for patients with moderately active ulcerative colitis. Gastroenterology 2009;137:1934-43.e1-3.
10. Feagan BG, Sandborn WJ, D'Haens G, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. Gastroenterology 2013;145:149-57.e2.
11. Chen P-J, Lin M-C, Lai M-J, et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 2018;154:568-75.
12. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94-100.
13. 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.
14. Hou JK, Tan M, Stidham RW, et al. Accuracy of diagnostic codes for identifying patients with ulcerative colitis and Crohn's disease in the Veterans Affairs Health Care System. Dig Dis Sci 2014;59:2406-10.
15. ImageNet. MATLAB toolbox for the ImageNet database. Available at: http://image-net.org/download-toolbox. Accessed October 15, 2019.
16. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV2016.:2818-26.
17. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014;1412:6980.
18. Yao H, Stidham RW, Soroushmehr R, et al. Automated detection of non-informative frames for colonoscopy through a combination of deep learning and feature extraction 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany 2019:2402-6.
19. Klein D. Implementing a general framework for assessing interrater agreement in Stata. Stata J 2018;18:871-901.
20. ASGE Technology Committee; Murad FM, Banerjee S, Barth BA, et al. Image management systems. Gastrointest Endosc 2014;79:15-22.
21. Ikeya K, Hanai H, Sugimoto K, et al. The Ulcerative Colitis Endoscopic Index of Severity more accurately reflects clinical outcomes and long-term prognosis than the Mayo Endoscopic Score. J Crohns Colitis 2016;10:286-95.
22. Takenaka K, Ohtsuka K, Fujii T, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020;158:2150-7.



Supplementary Figure 1. Learning curves for model performance based on the number of still images or full videos used for informative image and disease severity classifiers. **A**, Informative image learning curve. **B**, Disease severity learning curve. To build the learning curve for informative image and disease severity classifier images, the number of images from the high-resolution dataset used to train the classifier (cases, x axis) are plotted against the resulting area under the curve (AUC) using images from the external test set. The informative image classifier did not improve further beyond 25,000 still-image training cases. Further, the disease severity classifier did not substantially improve beyond 8000 training cases. Together, these data suggest adequate training set sample size.

SUPPLEMENTARY TABLE 1. Agreement on Mayo endoscopic scoring between 2 reviewers using the local developmental high-resolution video set

	Reviewer B			
	Mayo 0	**Mayo 1**	**Mayo 2**	**Mayo 3**
Reviewer A				
Mayo 0	18	2	0	0
Mayo 1	1	7	2	0
Mayo 2	0	1	6	1
Mayo 3	0	0	1	12

Two independent reviewers agreed on exact Mayo endoscopic score in 84.3% of high-resolution endoscopic videos reviewed ($\kappa = 0.95$). All disagreements were within 1 scoring level difference.