# Rapid development of accurate artificial intelligence scoring for colitis disease activity using applied data science techniques



Authors

**Mehul Patel[1], Shraddha Gulati[1], Fareed Iqbal[2], Bu'Hussain Hayee[1]**

Institutions

1 Department of Endoscopy, King's College Hospital NHS Foundation Trust, London

2 Surgease Innovations Ltd, London

**submitted** 23.8.2021

**accepted after revision** 14.12.2021

Bibliography

Endosc Int Open 2022; 10: E539–E543

**DOI** 10.1055/a-1790-6201

**ISSN** 2364-3722



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Corresponding author

Bu'Hussain Hayee, Department of Gastroenterology, 2nd Floor Hambleden Wing East, King's College Hospital, Denmark Hill, London SE5 9RS, UK

Fax: +442032996474

b.hayee@nhs.net

Supplementary material is available under https://doi.org/10.1055/a-1790-6201

## ABSTRACT

**Background and study aims** Scoring endoscopic disease activity in colitis represents a complex task for artificial intelligence (AI), but is seen as a worthwhile goal for clinical and research use cases. To date, development attempts have relied on large datasets, achieving reasonable results when comparing normal to active inflammation, but not when generating subscores for the Mayo Endoscopic Score (MES) or ulcerative colitis endoscopic index of severity (UCEIS).

**Patients and methods** Using a multi-task learning framework, with frame-by-frame analysis, we developed a machine-learning algorithm (MLA) for UCEIS trained on just 38,124 frames (73 patients with biopsy-proven ulcerative colitis). Scores generated by the MLA were compared to consensus scores from three independent human reviewers.

**Results** Accuracy and agreement (kappa) were calculated for the following differentiation tasks: (1) normal mucosa vs active inflammation (UCEIS 0 vs ≥1; accuracy 0.90, $\kappa = 0.90$); (2) mild inflammation vs moderate-severe (UCEIS 0–3 vs ≥4; accuracy 0.98, $\kappa = 0.96$); (3) generating total UCEIS score ($\kappa = 0.92$). Agreement for UCEIS subdomains was also high ($\kappa = 0.80$, 0.83 and 0.88 for vascular pattern, bleeding and erosions respectively).

**Conclusions** We have demonstrated that, using modified data science techniques and a relatively smaller datasets, it is possible to achieve high levels of accuracy and agreement with human reviewers (in some cases near-perfect), for AI in colitis scoring. Further work will focus on refining this technique, but we hope that it can be used in other tasks to facilitate faster development.

## Introduction

Endoscopy is endorsed by clinical guidelines as the most accurate method to stage disease activity in ulcerative colitis (UC) [1, 2], with meaningful correlates including long-term remission [3], risk of colectomy [4], weaning of steroids [5], and improved quality of life [6, 7]. Regulatory agencies mandate endoscopic evaluation in clinical trials, so accurate scoring also directly impacts development of new treatments [8, 9].

The four-category Mayo Endoscopic Sub-score (MES) has been credited for its ease of use, but remains unvalidated [10] and performs poorly when compared to the ulcerative colitis endoscopic index of severity (UCEIS) [4, 11]. The UCEIS, how-

ever, requires training and experience to be implemented properly and can take longer to perform than the MES.

Scoring colitis activity represents a complex task for artificial intelligence (AI). To date, studies of AI models based on the either MES [12–14] or UCEIS [15, 16] achieve reasonable accuracy for differentiating endoscopic remission from active disease, but performance for individual scores is less impressive. Furthermore, these developments required prohibitively large datasets.

Here we demonstrate that it is possible, by combining data science methods adapted for our purposes, to develop a highly accurate deep neural network for the complex task of UCEIS classification using a significantly smaller, but high-quality dataset.

## Methods

### Video capture and annotation

High-definition video recordings (MPEG-4, 1920×1080 at 25 frames per second) obtained from a prospective study (clinicaltrials.gov NCT04085211; LREC ref: 19/EM/0167) were used to develop a neural network for endoscopic scoring of UC. Fujifilm EC760 zoom-type colonoscopes were used throughout, restricted by the study intent. Video clips were extracted, unrestricted to anatomy, by a researcher blinded to patient details, disease extent or severity, managed and then scored using a previously-described methodology to prepare videos on a proprietary platform (Cord Vision, Cord Technologies, UK) [17].

### Endoscopic scoring

All video recordings were scored for the most inflamed region in the video clip by three independent reviewers with extensive experience using the UCEIS. If there was disagreement between reviewers for at least one domain of the UCEIS, the reviewers watched the recordings together to reach consensus. Subsequently, one reviewer evaluated each video recording on a frame-by-frame basis. Individual frames uninterpretable by a human (due to motion artifact, bowel preparation, glare) were excluded. We excluded 49,981 (51.9 %) frames, and the remainder scored for each UCEIS domain. Patient details are given in ▶ **Table 1**.

### Study design and model development

Video recordings from 73 procedures were available for model development; 55 video recordings (38,124 frames) were used to develop and train the initial classification model. From the outset, 18 recordings (8,277 frames) recordings were reserved for validation. After video preparation as above, we designed a multi-task learning framework [18] in which multiple objectives (in this case the individual sub-scores of the UCEIS) were trained simultaneously in a model, using a shared common architecture. A full description is included in Supplementary Methods. The UCEIS on a frame-by-frame basis was compared between the final model and annotations from human review to determine study endpoints. Scores from the model were able to be superimposed onto live video for read-outs and further comparisons (▶ **Fig. 1**), mirroring a potential real-time clinical application.

▶ **Table 1** Patient details.

| | Training Set (n = 55) | Test Set (n = 18) | |
|---|---|---|---|
| Age (median years, IQR) | 38.0 (19) | 32.0 (10) | $P = 0.24$ |
| Sex (male/female) | 27/28 | 15/3 | $P < 0.01$ |
| Montreal Classification | | | |
| ▪ E1 | 21.8 % (12) | 16.7 % (3) | $P = 1.00$ |
| ▪ E2 | 40.0 % (22) | 33.3 % (6) | $P = 0.58$ |
| ▪ E3 | 38.2 % (21) | 50.0 % (9) | $P = 0.46$ |
| Medications | | | |
| ▪ Oral mesalazine | 74.0 % (40) | 82.4 % (14) | $P = 0.53$ |
| ▪ Topical mesalazine | 9.3 % (5) | 11.8 % (2) | $P = 1.00$ |
| ▪ Immunomodulator | 18.5 % (10) | 23.5 % (4) | $P = 1.00$ |
| ▪ Anti-TNF | 7.4 % (4) | 11.8 % (2) | $P = 1.00$ |
| ▪ Anti-integrin | 9.3 % (5) | 17.6 % (3) | $P = 0.90$ |
| ▪ JAK-inhibitor | 1.9 % (1) | 11.8 % (2) | $P = 0.42$ |
| Simple Clinical Colitis Activity Index (median score, IQR) | 4.0 (5.8) | 3.0 (6.0) | $P = 0.85$ |

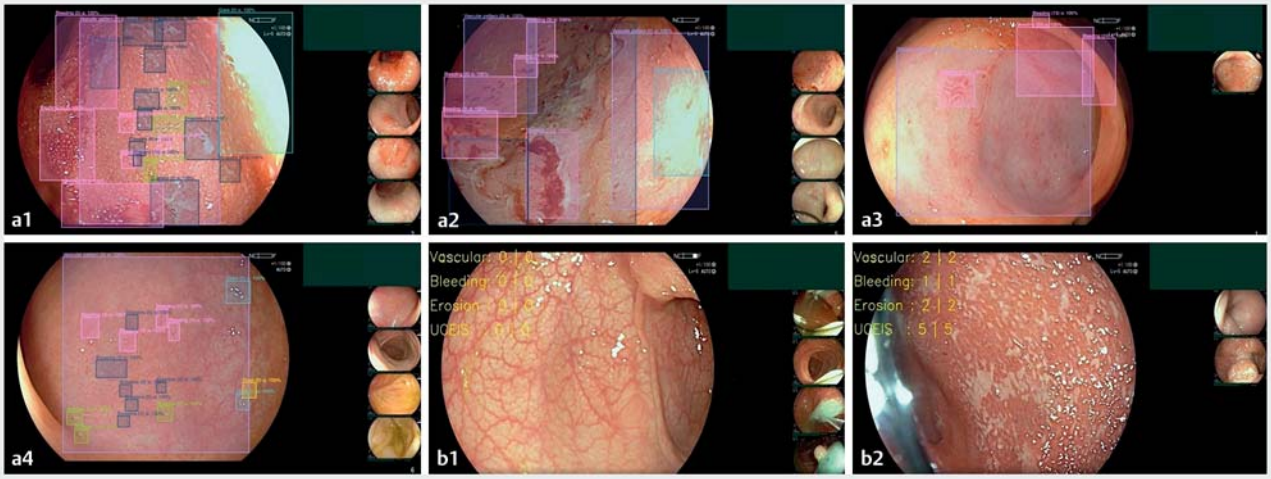
IQR, interquartile range.

### Study outcomes

The AI model was compared to the human consensus score on a frame-by-frame basis using the test set of videos. We evaluated model accuracy to: 1) distinguish endoscopic remission (UCEIS 0) from active disease; 2) distinguish mild (UCEIS 0–3) from moderate/severe disease; and 3) individual scores and subscores (▶ **Table 2**). The threshold for mild endoscopic disease activity is relevant to clinical practice with respect to treatment escalation. Secondary outcomes included agreement between: the model and expert human review for exact UCEIS scores; UCEIS domain scores (▶ **Table 3** and ▶ **Fig. 1**); and model performance to distinguish UCEIS 0/1 from > 1.

### Statistical analysis

Agreement between the final model and consensus human review was determined using quadratic weighted Cohen kappa coefficient (QWK) for both remission and distinguishing mild/moderate from severe disease. We generated confusion matrixes and subsequently determined sensitivity, specificity, positive predictive value and negative predictive value (NPV); 95 % confidence intervals were calculated. Statistical analysis was performed using RStudio version 1.3.959.




▶ **Fig. 1** Annotated videos were used in the development process (**a1–a4**) with multiple descriptors being tracked across video frames. The output from the final model, after training, was superimposed onto real-time video for the validation step (**b1, b2**) as might occur in a future clinical application.

▶ **Table 2** Summary of results for model performance on per frame analysis for distinguishing endoscopic remission and mild from moderate/severe disease.

| UCEIS study endpoint | Sensitivity | Specificity | PPV | NPV | Accuracy | QWK |
|---|---|---|---|---|---|---|
| 0 vs ≥ 1 | 0.93 (0.93–0.94) | 0.73 (0.71–0.76) | 0.95 (0.95–0.96) | 0.65 (0.62–0.67) | 0.90 (0.90–0.91) | 0.61 (0.61–0.65) |
| 0–3 vs ≥ 4 | 0.99 (0.99–1.00) | 0.98 (0.97–0.98) | 0.96 (0.96–0.97) | 0.99 (0.99–1.00) | 0.98 (0.98–0.98) | 0.96 (0.96–0.97) |

All results include 95 % confidence intervals in brackets.
UCEIS, ulcerative colitis endoscopic index of severity; NPV, negative predictive value; PPV, positive predictive value; QWK, quadratic weighted kappa statistic.

▶ **Table 3** Interobserver agreement for human reviewers (before consensus).

| | Fleiss Kappa | *P* value |
|---|---|---|
| Vascular pattern | 0.74 | *P*<0.001 |
| Bleeding | 0.76 | *P*<0.001 |
| Ulceration | 0.71 | *P*<0.001 |
| Total UCEIS | 0.75 | *P*<0.001 |

UCEIS, ulcerative colitis endoscopic index of severity.

▶ **Table 4** Intraclass correlation coefficient for UCEIS subdomains and total score between human scorers (after adjudication) and MLA.

| | Intraclass correlation coefficient (95 % CI) | *P* value |
|---|---|---|
| Vascular pattern | 0.81 (0.78–0.83) | *P*<0.001 |
| Bleeding | 0.71 (0.67–0.75) | *P*<0.001 |
| Ulceration | 0.88 (0.87–0.88) | *P*<0.001 |
| Total UCEIS | 0.92 (0.88–0.94) | *P*<0.001 |

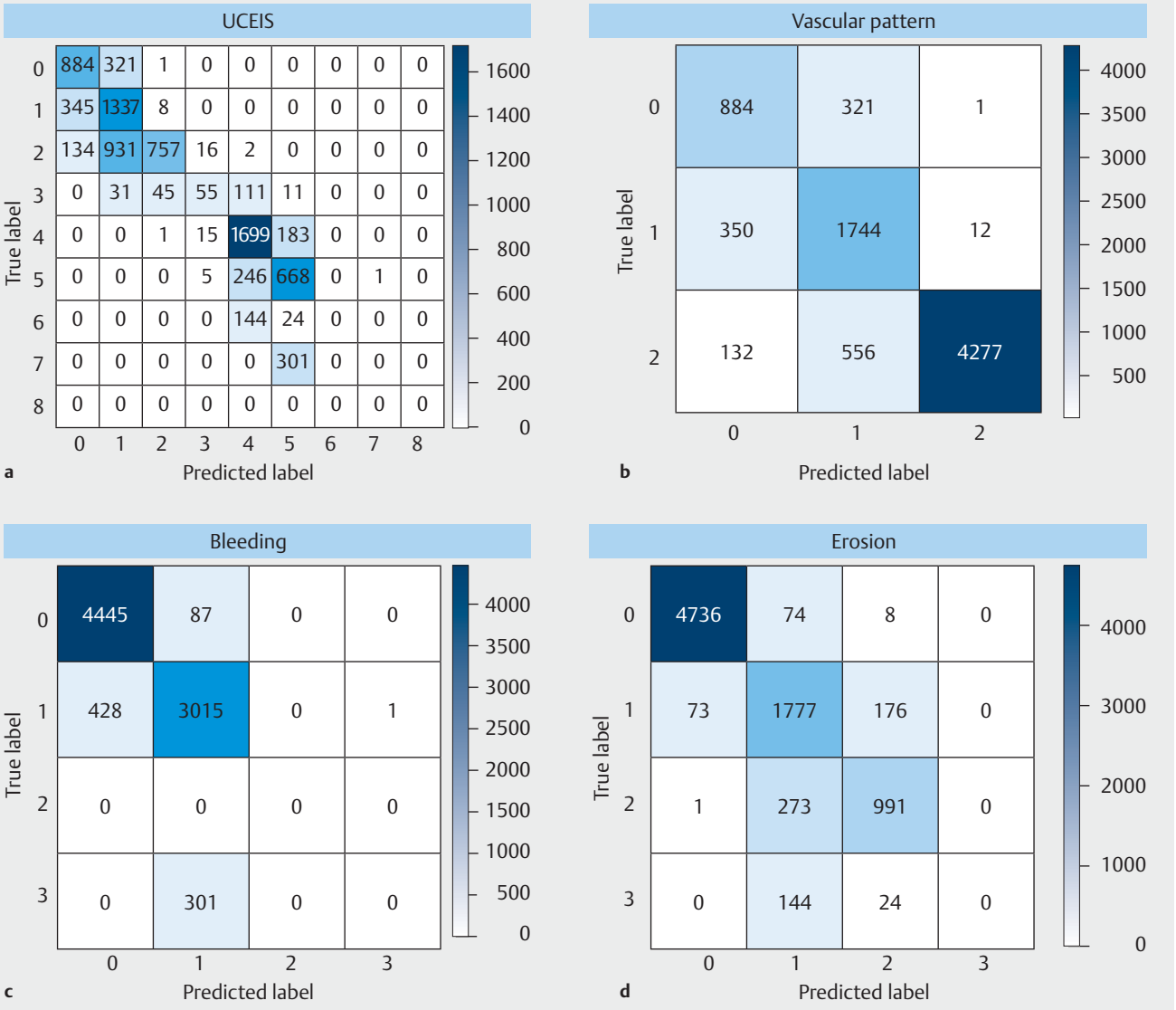
CI, confidence interval; MLA, machine learning algorithm; UCEIS, ulcerative colitis endoscopic index of severity

## Results

The model demonstrated a high level of accuracy across study endpoints (▶ **Table 2**). Additionally, there was near-perfect or substantial strength of agreement between the model and human review for UCEIS subdomains (▶ **Fig. 2** and ▶ **Table 4**). The interobserver agreement between human reviewers when scoring independently was less impressive, but still substantial (▶ **Table 3**).

## Conclusions

Our study shows it is possible to develop an accurate neural network for UCEIS scoring in a data efficient manner. We achieved this by performing per-frame scoring to maximize the value of data and selection of an appropriate model architecture. Accuracy was excellent for identifying endoscopic remission (UCEIS 0) and distinguishing mild from moderately active UC (UCEIS >3). There was an almost perfect agreement for




▶ **Fig. 2** Confusion matrixes comparing predicted model scores per frame against human review for test set across UCEIS domains. **a** Total UCEIS, **b** vascular pattern, **c** bleeding, **d** and erosion/ulceration.

the total UCEIS score and individual domains of the UCEIS between the model and human review on a per-frame basis. The use of only one endoscopy system to acquire video is a limitation of the study, but we will conduct further tests to establish accuracy across platforms (and will include training sets for other manufacturers). This does not limit the proof of concept that data science methods can reduce the burden of video acquisition.

Reliable and consistent UCEIS scoring remains a challenge in clinical practice. Validation studies for UCEIS scoring show that inter-investigator agreement for scoring is only moderate (k = 0.50). This was despite selecting investigators with an interest in inflammatory bowel disease (involvement in clinical trials) who undertook training sessions in UCEIS scoring [19]. It is likely that inter-investigator agreement is lower still in a real-world setting. An accurate AI model for colitis scoring has significant potential benefits stemming from standardized, consistent scoring without interobserver or intraobserver variation. Clinicians could reliably monitor a patient's endoscopic response to treatment over time; endoscopy would no longer be restricted to endoscopists with an interest in inflammatory bowel disease, therefore improving patient pathways; training tools for the novice endoscopist could be developed. In the context of therapeutic trials for UC, neural networks could remove the requirement for human central review of endoscopy recordings, reducing cost while improving confidence in reported trial outcomes.

There have been two other published studies evaluating models for UCEIS scoring. Gottlieb et al [16] used a larger dataset obtained from a multicenter drug trial involving 249 patients, 795 recordings, and 19.5 million frames. After an automated cleaning process, 61.5 % of frames were excluded, this

was higher than our study of 51.9 %. Unlike our study, they performed outcome analysis on a per-video recording basis for the overall UCEIS score. The definition of endoscopic healing was different to our study (UCEIS 0 vs 2–8, rather than 0 vs 1–8) and may, in part, explain the difference in accuracy for this task (97.04 % vs 90.0 %). This is also in the context of a much larger dataset which is inaccessible to the majority of researchers.

Performance of Gottlieb's model on an individual UCEIS score basis was excellent for UCEIS 0 (area under curve [AUC] = 0.885), but less impressive for the remaining scoring domains (e. g. AUC for UCEIS 1 = 0.333), which may explain the choice of definition for healing as above. Our use of a per-frame score instead of per-video score to train the model may have overcome this. In our study, when extending the definition of remission to UCEIS 0/1, accuracy was still high. Takenaka et al [15] used 40,758 still images, rather than video, from 875 patients to develop their model, after image cleaning, 4187 images were used to develop the model and 2000 images held back for a pilot study. Their model accuracy to predict endoscopic remission (90.1 %) was comparable to our study, but agreement between the model and human reviewers was lower (k = 0.80). Using still images may have limitations for extension into real-world, real-time applications.

We have shown our technique can accelerate the development of accurate models for even complex computer vision tasks with multiple parameters in one video sequence. Further validation can be conducted in real-world datasets to strengthen these observations; specifically at the extremes of the UCEIS score, but overall is expected to significantly shorten the time required to develop clinically useful and relevant models.

## Acknowledgments

This project was funded by a research grant from Surgease Innovations Ltd. Bu'Hussain Hayee is a minority shareholder in Surgease Innovations Ltd. None of the other contributors have any declarations. The authors would like to thank Eric Landau and Ulrik Hansen of Cord Technologies for data science support and Jabed Ahmed for video clip preparation and annotation.

### Competing interests

Dr. Hayee is a minority shareholder in Surgease Innovations Ltd.

### Funding

This research was supported by an unrestricted grant from Surgease Innovations Ltd.

### References

[1] Lamb CA, Kennedy NA, Raine T et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut 2019; 68: s1–106

[2] Rubin DT, Ananthakrishnan AN, Siegel CA et al. ACG Clinical Guideline: Ulcerative Colitis in Adults. Am J Gastroenterol 2019; 114: 384–413

[3] Turner D, Ricciuto A, Lewis A et al. STRIDE-II: An update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): Determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021; 160: 1570–1583

[4] Xie T, Zhang T, Ding C et al. Ulcerative Colitis Endoscopic Index of Severity (UCEIS) versus Mayo Endoscopic Score (MES) in guiding the need for colectomy in patients with acute severe colitis. Gastroenterol Rep 2018; 6: 38–44

[5] 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: 1194–1201

[6] Knowles SR, Keefer L, Wilding H et al. Quality of life in inflammatory bowel disease: a systematic review and meta-analyses – Part II. Inflamm Bowel Dis 2018; 24: 966–976

[7] Knowles SR, Graff LA, Wilding H et al. Quality of Life in Inflammatory Bowel Disease: A Systematic Review and Meta-analyses – Part i. Inflamm Bowel Dis 2018; 24: 742–751

[8] US 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

[9] European Medicines Agency. Guideline on the Development of New Medicinal Products for the Treatment of Crohn's Disease. Published Online First: 2019. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000375.jsp&mid=WC0b01ac0580032ec6%5Cnpapers2://publication/uuid/A052211A-E2B5-4084-9C22-29E99945A3C7

[10] Vashisht M, Samaan M, Mosli M et al. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Sytst Rev 2018; 1: CD011450

[11] 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 Crohn Colitis 2016; 10: 286–295

[12] Gutierrez Becker B, Arcadu F, Thalhammer A et al. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther Adv Gastrointest Endosc Artif 2021; 14: 1–15

[13] Bhambhvani HP, Zamora A. Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis. Eur J Gastroenterol Hepatol 2020: doi:10.1097/meg.0000000000001952

[14] 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 Net Open 2019; 2: e193963

[15] 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: 2150–2157

[16] Gottlieb K, Requa J, Karnes W et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology 2021; 160: 710–719.e2

[17] Hansen US, Landau E, Patel M et al. Novel artificial intelligence-driven software significantly shortens the time required for annotation in computer vision projects. Endosc Int Open 2021; 09: E621–E626

[18] Girshick R. Fast R-CNN. Proc IEEE Int Conf Comput Vis 2015: 1440–1448 doi:10.1109/ICCV.2015.169

[19] Travis SPL, Schnell D, Krzeski P et al. Reliability and initial validation of the ulcerative colitis endoscopic index of severity. Gastroenterology 2013; 145: 987–995