Transforming Ulcerative Colitis Care: AI-Powered Endoscopic Scoring from Clinical Trials to Clinical Practice

Authors:	*Flegg D,[1] Byrne MF[1-3] 1. University of British Columbia, Department of Internal Medicine, Vancouver, Canada 2. Dova Health Intelligence, Vancouver, Canada 3. Division of Gastroenterology, Vancouver General Hospital, University of British Columbia, Canada *Correspondence to df19@student.ubc.ca
Disclosure:	Byrne is a founder and shareholder in Dova Health Intelligence. Flegg has declared no conflicts of interest.
Received:	14.10.25
Accepted:	28.01.26
Keywords:	AI, automated endoscopic scoring, digital biomarkers, endoscopic disease activity, endoscopy, machine learning, ulcerative colitis (UC), video-based deep learning.
Citation:	EMJ. 2026;11[1]:36-40. https://doi.org/10.33590/emj/3LZ5MZ93

INTRODUCTION

Ulcerative colitis (UC) is a chronic, immune-mediated inflammatory disorder of the colon that follows a relapsing-remitting course. Characterised by mucosal inflammation, the disease requires accurate and reproducible monitoring tools to guide treatment, assess response, and evaluate remission. Endoscopic visualisation remains the gold standard for evaluating disease activity, with scores such as the Mayo Endoscopic Subscore (MES) and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) serving as core endpoints in both routine care and clinical trials.[1,2]

However, these indices are inherently subjective and prone to inter- and intra-observer variability, even among expert central readers.[3,4] Conventional scoring approaches often rely on assessing the most severely affected colonic segment, overlooking the heterogeneity of disease extent, and failing to capture spatial distribution. Patients with the same MES values may present with different endoscopic disease throughout their colon and significantly different clinical courses.[5]

The limitations of human scoring have known real-world consequences. They contribute to discordance between clinical symptoms and objective findings, delay therapeutic escalation, and introduce inefficiencies into clinical trials. Reader variability continues to be a key challenge in multicentre trials, where disagreements over endoscopic scores can lead to adjudication processes that slow recruitment and delay trial progression; moreover, such inconsistencies may misrepresent treatment response, complicating endpoint interpretation and trial outcomes.

In this context, AI has emerged as a promising solution to automate, standardise,

 Copyright © 2026 EMJ • CC BY-NC 4.0 Licence

and potentially enhance the assessment of endoscopic disease activity in UC. Through deep learning, a subset of AI that mimics the structure and function of the human brain, these systems can process vast volumes of imaging data and learn to recognise specific patterns or features within endoscopy videos. AI offers the potential to reduce reader variability, improve reproducibility, and support real-time decision-making, bringing objective and scalable tools into the hands of clinicians and researchers alike.

THE PROBLEM: VARIABILITY AND LIMITATIONS IN HUMAN SCORING

Despite their widespread use in both clinical practice and clinical trials, current endoscopic scoring systems in UC are plagued by substantial variability. Reproducibility is limited even among expert central readers, with key metrics such as MES and UCEIS showing only moderate agreement.[3] Discrepancies over critical thresholds, for example, differentiating MES 1 from MES 2, can directly influence patient eligibility for trials and assessments of therapeutic response.[4]

One of the core limitations of these indices lies in their design. Both MES and UCEIS are ordinal scales that typically assess only the worst-affected segment of the colon, overlooking the patchiness and extent of disease elsewhere. This segmental bias undermines their ability to reflect the true burden of inflammation and limits their sensitivity to therapeutic change.

In the context of clinical trials, such inconsistencies can introduce operational challenges. Differences in reader interpretation may necessitate adjudication processes, which can add logistical complexity and affect study timelines.[6] Subtle features such as endoscopic response and remission, particularly in moderate-to-severe disease or when mucosal healing is minimal, can be challenging to interpret and contribute to interobserver variability, even among experienced reviewers.[6,7] Such misclassifications may contribute to variability in treatment efficacy assessments, which in turn can influence the ability to meet primary endpoints and fully characterise therapeutic benefit.[8,9] In clinical practice, variability in scoring may contribute to differences in treatment decisions, including the potential for under-treatment of residual inflammation or overtreatment based on subjective interpretation.

These limitations have underscored the need for more reproducible and scalable solutions. As endoscopic endpoints become increasingly central to clinical trials, especially with treat-to-target strategies, the field must move beyond static, subjective indices. AI offers a compelling path forward by enabling standardised, segmental, and reproducible evaluation of disease severity, with the promise of transforming how UC is assessed across both clinical and research settings.[8]

ADVANCES IN AI-DRIVEN ENDOSCOPIC SCORING FOR ULCERATIVE COLITIS

AI has rapidly advanced the automation of endoscopic scoring in UC, evolving from early, static image-based classifiers to comprehensive video-based systems. Initial models, trained on thousands of annotated still frames, replicated expert assessments of disease severity but were limited by their reliance on curated, high-quality images and binary outcome predictions.[10,11] These approaches offered proof of concept but fell short in capturing the spatial and temporal complexity of colonic inflammation.

Recent developments have shifted towards dynamic, video-based models capable of analysing entire colonoscopy recordings. These systems preserve continuity across frames and allow for segmental evaluation, reflecting the heterogeneous nature of disease more accurately. Several groups have demonstrated strong correlation with central reader assessments using full-length videos, achieving high inter-rater reliability for endoscopic indices such as MES and UCEIS.[12-14]

CC BY-NC 4.0 Licence • Copyright © 2026 EMJ • March 2026 • EMJ

Byrne et al.[15] introduced a deep learning model capable of scoring UC activity under both MES and UCEIS at the frame-, section-, and video-level scales. Trained on over one million frames from full-length colonoscopies, the system showed concordance with expert readers. The quadratic weighted kappa used to compare the inter-rater agreement between expert's labels and the model's predictions showed strong agreement (0.87 and 0.88 at frame level; 0.88 and 0.90 at section level; and 0.90 and 0.78 at video level, for MES and UCEIS, respectively).[15]

Several recent studies have evaluated AI models for automated assessment of endoscopic disease activity in UC using full-length colonoscopy videos. Chaitanya et al.[16] described Arges (Janssen R&D, LLC, Raritan, New Jersey, USA), a transformer-based model that captures spatiotemporal patterns across full-length videos to estimate MES and UCEIS scores. Trained on millions of frames from four clinical trials and validated on held-out and prospective datasets (QUASAR), the agreement between the Arges MES output and human reader assessement (k=0.66) closely aligned with the two human expert rate agreement (k=0.71).[16] Gutierrez-Becker et al.[17] developed a model using spatial region mapping and per-frame MES classification, enabling segmental severity assessment and generating an interpretable Aggregate Disease Severity Score (ADSS).[17] Their performance measurements were based on internal validation, using held-out test sets from within their clinical trial data. There was high agreement between the model and central reading at the level of the colon section (k=0.80), and the agreement between central and local reading (k=0.84) suggested a similar inter-rater agreement between the model and experienced readers.

Stidham et al.[18] described the Cumulative Disease Score (CDS), a continuous metric derived from full-length video analysis, validated internally on held-out test sets. CDS was compared with the MES ($p<0.0001$) and all clinical components of the partial Mayo score ($p<0.0001$). CDS showed sensitivity to change, requiring 50% fewer participants to demonstrate endoscopic differences between ustekinumab and placebo than the MES.[18] More recently, Byrne et al.[19] developed a large-scale AI model trained on 83.6 million video frames that were labelled by three expert central readers across seven categories per video. For the MES, the Intraclass Correlation Coefficient (ICC) for three expert central readers was 0.905, and for the AI model and majority vote from the readers, the ICC was 0.907. The model also outputs continuous frame-by-frame level information on the sub-score characteristics that make up the UCEIS score. Collectively, these systems illustrate a broader shift towards video-based, segmental, and continuous AI-driven assessment of endoscopic disease activity. Rather than replicating human scoring alone, modern AI approaches aim to enhance reproducibility, reduce variability, and increase sensitivity in evaluating disease severity.

AI IN CLINICAL TRIALS AND REAL-WORLD PRACTICE

AI-based endoscopic scoring systems are increasingly being integrated into clinical trials for UC, where they offer improved reproducibility, eliminate reader bias, and streamline adjudication processes. These tools allow for more sensitive detection of treatment effects and facilitate more efficient trial designs by reducing sample size requirements.[19] By providing segmental, standardised scores across entire colonoscopy videos, AI systems are helping define more objective trial endpoints.

A recent example is the post-hoc application of AI scoring to the TITRATE study, a randomised trial comparing standard versus personalised guided infliximab dosing in acute severe UC. Although the original expert-read analysis did not meet the primary endpoint, AI-based re-analysis identified significantly higher response and remission rates in the personalised group.[9] While requiring prospective validation, these findings suggest that advanced video-based AI systems may reduce reader variability and

increase sensitivity in endpoint evaluation, highlighting their potential to refine clinical trial design.

In clinical practice, similar technologies are being adapted to support real-time disease monitoring and treat-to-target strategies. AI enables consistent scoring without reliance on central reading infrastructure, offering a scalable solution that supports both academic centres and community practices. When integrated into electronic health records or decision support platforms, these tools have the potential to enhance therapeutic decision-making and align care with evolving IBD management guidelines.

BARRIERS TO IMPLEMENTATION

Several barriers to widespread clinical implementation remain. Most AI systems have been developed using curated datasets with high-quality inputs, which may not reflect the variability encountered in routine practice. The reliance on expert-annotated training data is labour-intensive and limits scalability. Additionally, model performance often declines with suboptimal video quality, poor bowel preparation, or varying equipment, which highlights the need for robust validation across diverse clinical settings.[4,8] A further limitation lies in differentiating UC from other forms of colitis, such as infectious, ischaemic, Crohn's, or drug-induced colitis. Most AI tools are optimised for activity scoring rather than diagnostic classification, which may reduce their performance in nuanced or atypical cases. Finally, integration into real-world workflows requires solving logistical and regulatory hurdles related to real-time processing, interface design, and interpretability. Regulatory approval will be essential, as AI tools must meet safety, accuracy, and transparency standards set by regulatory bodies before widespread clinical adoption.

CONCLUSION

AI has rapidly evolved as a promising tool for standardising endoscopic assessment in UC. From early, static image-based models through convolutional networks, to advanced video-based systems with transformers and self-supervised learning, AI has learned from unlabelled data and now aligns closely with expert central readers. These advancements are reshaping how endoscopic data are interpreted, enabling more reproducible and granular evaluations than traditional indices like MES or UCEIS. Yet, these advancements do not replace human expertise. Hybrid human-AI models will be essential, ensuring AI supports clinical decision-making rather than replacing it.

Looking ahead, as more advanced and validated tools become more integrated into clinical infrastructure, AI is poised not only to transform clinical trials but also to influence everyday practice. Future directions include prospective validation trials, regulatory qualification for clinical use, and seamless integration into practice. Real-time, AI-supported scoring could offer consistent, objective evaluations across practice settings, supporting treat-to-target strategies and enabling more personalised, data-driven care for patients with UC.

References

1. Schroeder KW et al. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. N Engl J Med. 1987;317(26):1625-9.
2. Travis SP et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut. 2012;61(4):535-42.
3. Mohammed Vashist N et al. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Syst Rev. 2018;1(1):CD011450.
4. Murino A, Rimondi A. Automated artificial intelligence scoring systems for the endoscopic assessment of ulcerative colitis: how far are we from clinical application? Gastrointest Endosc. 2023;97(2):347-9.
5. Kim B et al. Endoscopic and histological patchiness in treated ulcerative colitis. Am J Gastroenterol. 1999;94(11):3258-62.
6. Hashash JG et al. Inter- and intraobserver variability on endoscopic scoring systems in Crohn's disease and ulcerative colitis: a systematic review and meta-analysis. Inflamm Bowel Dis. 2024;30(11):2217-26.
7. Osada T et al. Comparison of several activity indices for the evaluation of

CC BY-NC 4.0 Licence • Copyright © 2026 EMJ

endoscopic activity in UC: inter- and intraobserver consistency. Inflamm Bowel Dis. 2010;16(2):192-7.

8. Lee MCM et al. Artificial intelligence for classification of endoscopic severity of inflammatory bowel disease: a systematic review and critical appraisal. Inflamm Bowel Dis. 2025;31(8):2296-310.
9. Gecse KB et al. DOP098 AI tool distinguishes differences in endoscopic disease activity in ulcerative colitis where humans could not: data from the TITRATE trial. J Crohn's Colitis. 2025;20(Suppl 1):jjaf231.135.
10. Stidham RW 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(5):e193963.
11. Takenaka K et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology. 2020;158(8):2150-7.
12. Gottlieb K et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology. 2021;160(3):710-9.e2.
13. Yao H et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc. 2021;93(3):728-36.e1.
14. Fan Y et al. Novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. Gastrointest Endosc. 2023;97(2):335-46.
15. Byrne MF et al. Application of deep learning models to improve ulcerative colitis endoscopic disease activity scoring under multiple scoring systems. J Crohns Colitis. 2023;17(4):463-71.
16. Chaitanya K et al. Arges: spatio-temporal transformer for ulcerative colitis severity assessment in endoscopy videos. arXiv. 2024;DOI:10.48550/arXiv.2410.00536.
17. Gutierrez-Becker B et al. Ulcerative colitis severity classification and localized extent (UC-SCALE): an artificial intelligence scoring system for a spatial assessment of disease severity in ulcerative colitis. J Crohns Colitis. 2025;19(1):jjae187.
18. Stidham RW et al. Using computer vision to improve endoscopic disease quantification in therapeutic clinical trials of ulcerative colitis. Gastroenterology. 2024;166(1):155-67.e2.
19. Byrne M et al. Development and validation of a novel AI-based computer vision solution for ulcerative colitis severity scoring on video for real world using high-volume expert-annotated video frames. UEG Journal. 2025;13(Suppl 8):704.

FOR REPRINT QUERIES PLEASE CONTACT: INFO@EMJREVIEWS.COM