ARGES-Ulcer: A high-performance, generalizable AI model for Ulcer Segmentation in Crohn's Disease from endoscopy videos

AUTHORS: Krishna Chaitanya[1], Fabio Gunderson[2], **Lukas Hirsch[2]**, Pooya Mobadersany[2], Chaitanya Parmar[2], Brendon Lutnick[2], Shinobu Yamamoto[2], Yuliya Halchenko[2], Nicholas Skomrock[2], Shadi Yarandi[2], Lindsey Surace[2], Louis R Ghanem[2], Michael D Rice[3], Tadd Hiatt[3], Ryan Stidham[3], Tommaso Mansi[2], Gabriela Oana Cula[2], Kristopher Standish[2], Pablo F Damasceno[1]. AFFILIATIONS: [1] Johnson & Johnson, Switzerland, [2] Johnson & Johnson, USA, [3] University of Michigan Medicine, Ann Arbor, MI, United States.

DDW2026 Digestive Disease Week® MAY 2-5, 2026 | CHICAGO, IL EXHIBIT DATES: MAY 3-5, 2026

Background

- The Simple Endoscopic Score for Crohn's Disease (SES-CD) is a widely used standard for assessing endoscopic disease severity in Crohn's disease (CD) clinical trials.
- Ulceration within SES-CD is graded categorically based on visual estimation of ulcerated mucosal surface area, leading to subjectivity, inter-reader variability, and limited sensitivity to change.
- These limitations can reduce reproducibility and obscure subtle but clinically meaningful differences in disease activity.
- ARGES-Ulcer is a computer-vision-based AI tool that performs pixel-level segmentation of ulcerated regions in endoscopy videos, enabling objective and reproducible ulcer burden quantification.
- By leveraging pixel-level outputs, ARGES-Ulcer enables continuous ulcer burden metrics that provide higher-resolution assessment than traditional ordinal SES-CD subscores.

Objective

- To validate ARGES-Ulcer analytically and clinically by assessing its generalizability for ulcer segmentation across multicenter trials and the clinical relevance of derived continuous ulcer burden scores.

Methods

- ARGES-Ulcer was trained using an encoder-decoder neural network to generate pixel-level ulcer masks based on human annotation.
- 6,114 expert-annotated frames from the SEAVUE Phase 3b [3]Crohn's disease trial were split into training/validation/holdout sets (60%/20%/20%),

I) **Analytical Validation** (Ulcer Segmentation):
- Model performance was evaluated on i) a SEAVUE [3] holdout set, and ii) an independent external dataset from TRIDENT Phase 2b [4] trial.
- To assess inter-rater variability, five expert gastroenterologists provided additional annotations on a separate set of 600 SEAVUE frame set.
- Segmentation performance was measured using Dice similarity coefficient (Dice) and Hausdorff Distance (HD), comparing AI-to-expert and expert-to-expert agreement.

II) **Clinical Validation** (Continuous Ulcer Burden Scoring):
- For each endoscopy video, pixel-level ulcer masks were used to compute frame-level ulcer percentage, which were averaged across video to derive a continuous ARGES-Ulcer$_{\text{percentage-score}}$ (0-100%).
- The model was applied to the GALAXI Phase 3 [5] CD trial at Screening, Week 12, and Week 48.
- Associations with endoscopic, clinical, biomarker, histologic, and treatment-response outcomes were evaluated using Spearman correlation and Mann-Whitney U tests.

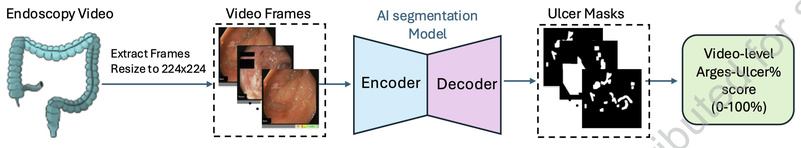

Figure 1: Automated estimation of ulcerated surface from colonoscopy video. Individual frames from endoscopy videos were processed by ARGES-Ulcer model to generate ulcer segmentation masks. Frame-level ulcer percentage was calculated as the proportion of ulcer pixels relative to total pixels. The video-level ARGES-Ulcer percentage score was defined as the mean of frame-level percentages across the entire video, yielding a continuous score from 0 to 100%.

https://www.congresshub.com/Immunology/DDW2026/Guselkumab/
Chaitanya-ARGES-Ulcer

This QR code is intended to provide scientific information for individual reference and the information should not be altered or reproduced in any way.

Results

I) Analytical Validation (Ulcer Segmentation):

a) ARGES-Ulcer demonstrated consistent ulcer segmentation performance across two independent multicenter clinical trials (SEAVUE 3] and TRIDENT[4]) and comparable to published IBD ulcer segmentation models [1,2].

Metric	SEAVUE (Hold-out)	TRIDENT
Dice (↑)	0.626+/-0.26	0.552+/-0.19
HD (↓)	45.07+/-37.12	56.03+/-28.19

Table 1: Performance metrics: Dice (higher is better indicated by up arrow (↑)) and Hausdorff Distance (HD; lower is better indicated by down arrow (↓)) for AI-based ulcer segmentation overlap with ground-truth annotations across two clinical trial datasets.

b) AI-to-expert agreement is comparable to expert-to-expert variability.

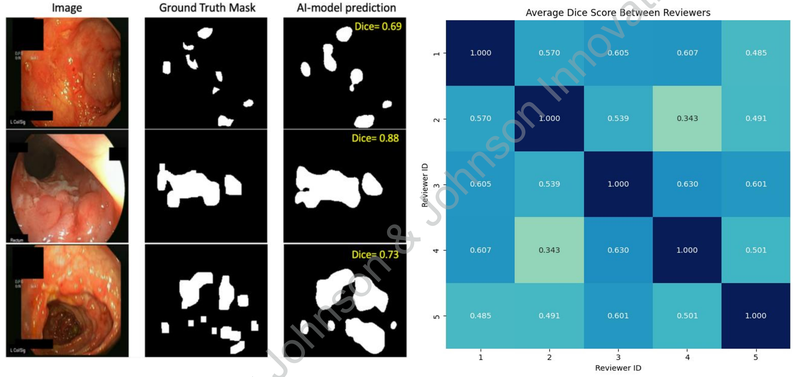

Figure 2: Left: Visual comparison of blinded, expert ground-truth annotations vs AI-estimated ulcer masks in frames from different anatomical locations. **Right:** Inter-rater variability among ulcer segmentations from five experts over 600 frames.

II) Clinical Validation (ARGES-Ulcer$_{\text{percentage-score}}$: Continuous Ulcer Burden)

a) Changes in ARGES-Ulcer$_{\text{percentage-score}}$ shows strong, significant associations with changes in many disease activity markers for GALAXI [5].

Clinical Measures	ARGES-Ulcer$_{\text{percentage-score}}$	
Diease Severity scores	Correlation coefficient	p-value
SES-CD and subscores	0.56	<0.001
Ulcerated surface area	0.54	<0.001
Affected surface area	0.56	<0.001
Presence and Size of Ulcers	0.53	<0.001
Narrowing	0.05	0.1060
CDAI score	0.30	<0.001
Biomarkers		
Fecal calprotectin (FCP)	0.32	<0.001
C-reactive protein (CRP)	0.46	<0.001
Histological Score		
Total GHAS Score	0.49	<0.001

Table 2: Association between ARGES-Ulcer$_{\text{percentage-score}}$ and disease severity, biomarker and histological scores on hold-out GALAXI CD cohort. Reported values are Spearman correlation coefficients and two-tailed p-values.

Johnson&Johnson

b) Lower ARGES-Ulcer$_{\text{percentage-score}}$ is associated with clinical remission at Week 48.

c) Greater reductions in ARGES-Ulcer$_{\text{percentage-score}}$ are associated with treatment

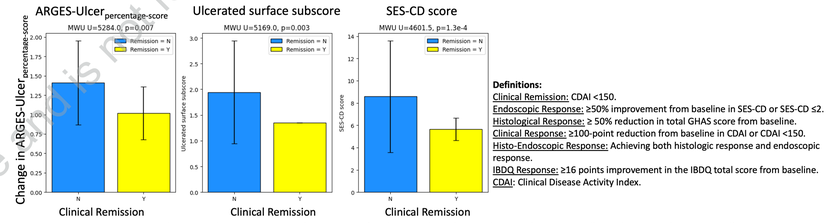

Figure 3A: Association of Arges-Ulcer$_{\text{percentage-score}}$ (continuous), Ulcerated surface subscore (categorical), and SES-CD (categorical) scores with clinical remission at Week 48 on the prospective GALAXI data. Subjects achieving remission show lower scores compared to non-remitters.

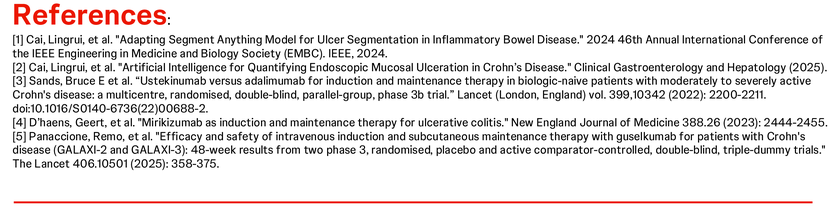

Figure 3B: Association between change in Arges-Ulcer$_{\text{percentage-score}}$ (Week 48 vs. Screening) and treatment response variables: Remission, Endoscopic response, Histo-endoscopic response, Histological response, and IBDQ response (p-values<=0.01) on the prospective GALAXI data.

Discussion

- ARGES-Ulcer demonstrated robust and generalizable segmentation performance across internal and external multicenter clinical trial datasets.
- AI segmentation accuracy was comparable to published IBD ulcer segmentation models[1,2].
- Substantial inter-rater variability among expert readers highlights the subjectivity of manual ulcer assessment.
- AI-to-expert agreement matched expert-to-expert consistency, supporting reliable and reproducible performance across readers and trials.
- The derived continuous ARGES-Ulcer$_{\text{percentage-score}}$ showed clinically meaningful associations with endoscopic, clinical, biomarker, and histologic measures, supporting its added value over categorical scoring.

Conclusion

- ARGES-Ulcer achieves expert-level performance for pixel-level ulcer segmentation on unseen, multicenter Crohn's disease clinical trial data.
- AI-derived continuous ulcer burden score is clinically meaningful, showing strong associations with endoscopic, clinical, biomarker, histologic, and treatment-response endpoints.
- This enables rapid, consistent, and sensitive endoscopic assessment, addressing limitations of categorical scoring & support more reproducible disease severity evaluation in CD clinical trials.

References:

[1] Cai, Lingrui, et al. "Adapting Segment Anything Model for Ulcer Segmentation in Inflammatory Bowel Disease." 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2024.
[2] Cai, Lingrui, et al. "Artificial Intelligence for Quantifying Endoscopic Mucosal Ulceration in Crohn's Disease." Clinical Gastroenterology and Hepatology (2025).
[3] Sands, Bruce E et al. "Ustekinumab versus adalimumab for induction and maintenance therapy in biologic-naive patients with moderately to severely active Crohn's disease: a multicentre, randomised, double-blind, parallel-group, phase 3b trial." Lancet (London, England) vol. 399,10342 (2022): 2200-2211. doi:10.1016/S0140-6736(22)00688-2.
[4] D'haens, Geert, et al. "Mirikizumab as induction and maintenance therapy for ulcerative colitis." New England Journal of Medicine 388.26 (2023): 2444-2455.
[5] Panaccione, Remo, et al. "Efficacy and safety of intravenous induction and subcutaneous maintenance therapy with guselkumab for patients with Crohn's disease (GALAXI-2 and GALAXI-3): 48-week results from two phase 3, randomised, placebo and active comparator-controlled, double-blind, triple-dummy trials." The Lancet 406.10501 (2025): 358-375.

Acknowledgements & Disclosures

We would like to thank Tadeusz Lewandowski, Jie Wang, Tamsin Sargood, Lourdes Zellhofer, Aleksandar Stojmirovic for assistance with data preparation. This work was supported by Janssen Research & Development, LLC.

All development of AI models and analyses presented in this poster has been conducted internally within Janssen. K. Chaitanya, F. Gunderson, L. Hirsch, P. Mobadersany, C. Parmar, B. Lutnick, S. Yamamoto, Y. Halchenko, N. Skomrock, S. Yarandi, L. Surace, L.R. Ghanem, T. Mansi, G.O. Cula, K. Standish, P. Damasceno are employees of Johnson & Johnson and own company stock/stock options.

CONTACT: Pablo Damasceno (pdamasc1@its.jnj.com)