Machine learning assessment of endoscopic severity in Ulcerative Colitis trials: model evaluation against the 2+1 reference standard (DOP078)

Chakib Battioui[1], Pavel Brodskiy[2], Klaus Gottlieb[1], Mohammad Haft-Javaherian[2], William Eastman[1], Julian Lehrer[2], Evan Yu[2], Derek Onken[1], Darren Thomason[2], Daniel Colucci[2], Walter Reinisch[3], Shrujal Baxi[2]

[1]Eli Lilly and Company, Indianapolis, IN, USA

[2]Iterative Health Inc., Cambridge, MA, USA

[3]Medical University of Vienna, Vienna, Austria

Conflict of Interest Disclosures

CB, KG, WE, and **DO** are employees of Eli Lilly & Co. and may hold stock or stock options in Eli Lilly & Co.

PB, MH, JL, EY, DT, DC, and **SB** are employees of Iterative Health, Inc. and may hold stock or stock options in Iterative Health, Inc.

WR has served as a speaker for AbbVie, Celltrion, Ferring, Janssen, Galapagos Medice, MSD, Roche, Pfizer, Sobi, Takeda, as a consultant for AbbVie, Amgen, AOP Orphan, Boehringer Ingelheim, Bristol Myers Squibb, Calyx, Celltrion, Eli Lilly, Galapagos, Gilead, Index Pharma, Janssen, Medahead, Microbiotica, Pfizer, Teva, Takeda; as an advisory board member for AbbVie, Amgen, Boehringer Ingelheim, Bristol Myers Squibb, Celltrion, Galapagos, Janssen, Pfizer, Teva and has received research funding from AbbVie, Janssen, Sandoz, Sanofi, Takeda.

Background

- Regulatory guidance recommends the endoscopy subscore as a primary endpoint in ulcerative colitis (UC) therapeutic trials.
  - Typically assessed via the 2+1 central reading paradigm.
- High variability exists among expert assessments of the endoscopy subscore, impacting the reliability and reproducibility of trial results.
- Machine learning (ML) provides an opportunity for standardization.

Objective: Evaluate a novel machine learning model to assess the endoscopy subscore on complete endoscopy procedure recordings from a UC clinical trial compared to a 2+1 reference standard.

Methods

- Videos from mirikizumab Phase 2 (NCT02589665) and Phase 3 induction (NCT03518086) trials in UC were added to database of endoscopic recordings from routine practice.
    - 18,169 total videos in study cohort.
- 639 videos (~25%) from Phase 3 induction trial (week 0 and 12 procedures) with a 2+1 centrally read endoscopy subscore were randomly selected and held out to evaluate performance of the final, locked model.
    - Holdout test dataset isolated from the development team.
    - Distribution of severity similar to study population (82.2% moderate-to-severe on endoscopy).
    - 62.4% agreement rate between the first two human readers (local vs central reader 1) in determining the 2+1 endoscopic subscore, in line with published data.
- Remaining videos used to develop a state of the art, multi-stage, deep learning algorithm to assess the endoscopy subscore on UC videos.
- Quadratic weighted kappa (QWK) was used to evaluate inter-rater agreement between the ML model assessed endoscopy subscore and 2+1 reference standard on the complete endoscopy procedure recordings.

Results

Assessment	Performance Metric	Value
Ordinal endoscopic subscore (0, 1, 2, 3)	QWK	0.77
	Accuracy	70.0%
	Specificity	89.1%
	Sensitivity	62.6%
Endoscopic improvement (0, 1 vs 2, 3)	Accuracy	89.8%
Endoscopic remission (0 vs 1, 2, 3)	Accuracy	94.2%

Table 1. Top-line performance metrics comparing the ML model to the 2+1 centrally read reference standard endoscopy subscore on complete endoscopy recordings. Abbreviations: QWK, quadratic weighted kappa.

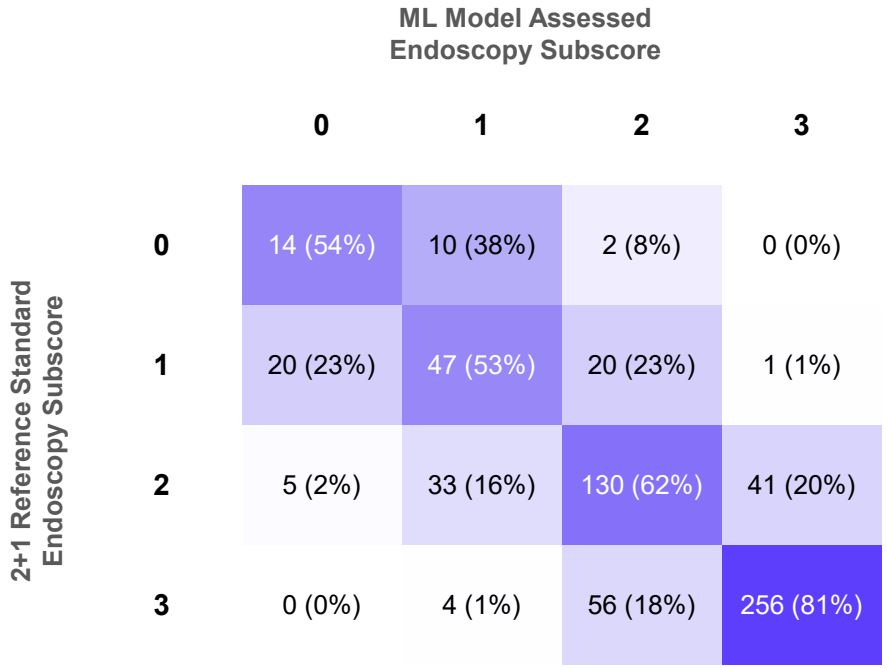

Figure 1. 4x4 confusion matrix comparing the ML model to the 2+1 centrally read reference standard endoscopy subscore on complete endoscopy recordings.

Conclusion

This ML algorithm effectively assesses the endoscopy subscore on full-length endoscopic videos in UC.

- Given the performance, this ML model could standardize endoscopic disease activity assessments in prospective trials.
  - Addressing a notable challenge currently posed in UC therapeutic development programs due to inconsistencies among human readers.
- Future research will investigate ML-based reading paradigms for the assessment of endoscopic endpoints in trials.
  - We invite you to view our related posters investigating ML-based multi-reader paradigms (P0307 and P0430).

Presentation & Lilly Content QR Codes

Scan QR code to access this **presentation**

Scan QR code for **a list of all Lilly content** presented at ECCO 2025. Other company and product names are trademarks of their respective owners.