

Interobserver and intra-observer variability in bowel preparation scoring for colon capsule endoscopy: impact of AI-assisted assessment, interim analysis

Ian Io Lei¹, Daniel Gaya², Alexander Robertson³, Benedicte Schelde-Olesen⁴, Alice Mapiye³, Anirudh Bhandare⁵, Ursula Valentiner⁶, Pablo Laiz⁷, Santi Segui⁸, Anastasios Koulaouzidis⁹, Ramesh Arasaradnam¹

¹University Hospital Coventry and Warwickshire

²University of Glasgow Medical School

³University Hospitals of Leicester NHS Trust

⁴Odense University Hospital

⁵Royal Oldham Hospital

⁶Institute of Anatomy and Experimental Morphology, University Medical Center Hamburg-Eppendorf

⁷Corporate Health International

⁸University of Barcelona

⁹Odense University Hospital

Introduction

Colon capsule endoscopy (CCE) has gained prominence post-coronavirus 2019 (COVID-19) as a non-invasive alternative for lower gastrointestinal investigations. However, bowel cleansing remains challenging, because CCE cannot suction, wash or reposition for better mucosal visualisation. While interobserver variability in bowel preparation scoring is well documented in traditional colonoscopy, its impact on whole-video CCE assessment remains unclear.

Objective

This CESCAIL sub-study aimed to evaluate interobserver agreement in CCE bowel cleansing assessment among readers and assess the agreement between AI-assisted and manual assessments and whether AI improves interobserver agreement.

Material and Methods

As part of the CESCAIL study, 25 completed videos were randomly selected from 673 CCE recordings. Nine readers with varying levels of CCE experience assessed bowel cleansing quality using the Leighton Rex scale and the Colon Capsule CLEansing Assessment and Report (CC-CLEAR) score. Following a 6-month washout period, the same readers reassessed the videos using AI-assisted analyses through CC-CLEAR score to evaluate improvements in interobserver variability and changes in intra-observer variability between manual and AI-assisted readings. Interobserver variability was assessed using intraclass correlation coefficients (ICC) and bootstrapping with 1,000 iterations with Fleiss Kappa agreement, while intraobserver variability was evaluated using Cohen Kappa agreement and Bland-Altman analysis.

Result and Discussion

The Leighton Rex scale showed moderate reliability (ICC=0.55, 95% CI: 0.48–0.63), while CC-CLEAR demonstrated good reliability (ICC=0.89, 95% CI: 0.86–0.92), significantly reducing interobserver variability. Clinician agreement was poor ($\kappa=0.0889$), but AI-assisted scoring improved it to moderate levels ($\kappa=0.3419$, $p=0.0098$). The intraobserver agreement between AI-assisted and manual assessment showed moderate to excellent reliability (ICC=0.69–0.90). Cohen's Kappa analysis revealed good agreement between manual and AI-assisted evaluation among experienced readers ($\kappa=0.67$ –0.85) but moderate agreement among less experienced readers ($\kappa=0.47$ –0.61). However, Bland-Altman analysis showed that AI-assisted assessment consistently assigned lower bowel cleansing scores compared with manual reading.

Conclusion

Interobserver agreement for CC-CLEAR was good, irrespective of the readers' experience levels. AI-assisted assessments significantly improved interobserver agreement, providing more consistent and reproducible scoring. The moderate agreement between AI-assisted and manual assessments suggests that further optimisation is needed to enhance alignment with manual scoring while preserving the strong interobserver agreement. Full study results are forthcoming.

doi: [10.1016/j.clinme.2025.100421](https://doi.org/10.1016/j.clinme.2025.100421)