Diseases of the Esophagus (2026), **39**, 1–7
https://doi.org/10.1093/dote/doaf133

DISEASES OF THE ESOPHAGUS

ISDE The International Society for Diseases of the Esophagus

Original Article

Interobserver agreement for the assessment of erosive reflux esophagitis: a *post hoc* analysis of clinical trial data

Dennis Wang,[1] Kayla Dadgar,[2] Yuhong Yuan,[1] Paul Sinclair,[3] Prateek Sharma,[4] Michael Vaezi,[5] David Armstrong,[1,6,*]

[1]*Division of Gastroenterology, McMaster University, Hamilton, ON, Canada,* [2]*Division of Gastroenterology, University of Alberta, Edmonton, ON, Canada,* [3]*INSINC Consulting Inc., Guelph, ON, Canada,* [4]*Department of Gastroenterology, VA Kansas City and the University of Kansas Medical Center, Kansas City, MO, USA,* [5]*Director, Center for Swallowing and Esophageal Disorders, Vanderbilt University Medical Center, Nashville, TN, USA, and* [6]*Farncombe Family Digestive Health Research Institute, McMaster University, Hamilton, ON, Canada*

SUMMARY. **Interobserver agreement for the Los Angeles (LA) classification of erosive reflux esophagitis was good in validation studies, but limited agreement data exists from clinical trials (CTs). We conducted a *post hoc* evaluation of interobserver agreement between CT endoscopists and independent expert adjudicators in a multi-center, randomized controlled trial of a new acid suppression therapy. Trial endoscopists captured endoscopic images/videos and documented esophagitis severity using the LA classification. Adjudicators reviewed images/videos on a web-based platform. If the first two adjudicators disagreed and the third adjudicator did not produce a majority verdict, all three conferred to reach consensus. Cohen's kappa (κ) evaluated interobserver agreement. Cohen's weighted kappa (κ_w) evaluated agreement corrected for disagreement extent. Of 388 images/videos with adequate quality, trial endoscopists and adjudicators agreed on esophagitis severity in 168 (43.3%) cases, and assigned more severe grades than adjudicators for 185 (47.7%) cases. Agreement was fair between trial endoscopists and adjudicators (κ: 0.27; κ_w: 0.40), moderate between individual adjudicators (κ: 0.43 to 0.47), and good between adjudicators and final diagnosis (κ: 0.75 to 0.78). After adjusting for disagreement extent, agreement was good between individual adjudicators (κ_w: 0.63 to 0.66), and very good between adjudicators and final diagnosis (κ_w: 0.84 to 0.87). Interobserver agreement on esophagitis severity between CT endoscopists and adjudicators was fair. Initial agreement between adjudicators was moderate, but agreement between adjudicators and consensus diagnosis was very good. Accurate esophagitis grading for CTs requires further training on LA classification and a robust central reading protocol.**

KEY WORDS: **erosive reflux esophagitis (ERE), gastroesophageal reflux disease (GERD), acid reflux, gastrointestinal endoscopy, clinical trials.**

Address correspondence to: David Armstrong, Douglas Family Chair in Nutrition Research, Farncombe Family Digestive Health Research Institute & Division of Gastroenterology, McMaster University Medical Centre 1280 Main Street West, HSC-3V41, Hamilton, Ontario L8S 4K1 Canada. Email: armstro@mcmaster.ca

Conflicts of interest: Dennis Wang: no disclosures. Kayla Dadgar: no disclosures. Yuhong Yuan: received fees for data analysis from IWGCO. Paul Sinclair: received fees for data analysis from IWGCO. Michael Vaezi: board member and consultant for IWGCO; consultant for Diversatek, Phathom Pharmaceuticals, Isothrive, Sanofi. Prateek Sharma: president and consultant for IWGCO; consultant for Olympus Corporation, Boston Scientific, Salix Pharmaceuticals, Cipla, Takeda, Samsung Bioepis, CDx, Exact Science; receives research funding from Fujifilm, ERBE Medical, Ironwood Pharmaceuticals, and NEC. David Armstrong: Co-founder & Chief Innovation Officer: A.I. VALI Inc.; advisory boards for Cinclus Pharma, Phathom Pharmaceuticals, Canadian Digestive Health Foundation. Sanofi (remunerated); speakers' fees for Ferring, Takeda; Treasurer for IWGCO; receives research funding from Nestlé Health Sciences, Weston Family Foundation. David Armstrong was supported by the Douglas Family Chair in Nutrition Research. We appreciate the unrestricted funding from Cinclus Pharma to analyze and report these data.
Financial support: None declared.

INTRODUCTION

Reflux esophagitis affects many people worldwide, with prevalence in the general population of 3.4% to 15.6%.[1–3] The Los Angeles (LA) classification for the endoscopic assessment of reflux esophagitis was first introduced in 1994.[4] Erosive reflux esophagitis (ERE) was defined by mucosal breaks, with the severity of esophagitis ranging from A to D based on the length and circumferential extent of these breaks. The classification was developed to create a standardized, validated grading system to document ERE severity in an accurate, specific, and reproducible manner.[4]

Since its inception, the LA classification has become widely adopted in clinical practice and research. Various guidelines have incorporated the LA classification in the diagnosis and management of gastroesophageal reflux disease.[5,6] Esophagitis severity based on the LA classification has been associated with other clinical factors, such as esophageal acid exposure, daytime and night-time reflux episodes, and heartburn severity.[7–9] Patients with severe ERE (LA grades C and D) are less likely to achieve healing with medical therapy[10] and they frequently require long-term proton pump inhibitor (PPI) maintenance therapy as relapse rates are higher than for LA grade A or B esophagitis.[5]

Validation studies have reported good interobserver agreement for the LA classification, as well as similar or better interobserver agreement compared to other classification systems for reflux esophagitis.[4,7,11–14] However, there are limited data on interobserver agreement in grading esophagitis severity in clinical trials (CTs). In addition, there are a few studies in which central adjudicators grade ERE and have their assessments compared to local investigators.[15] Conversely, central adjudication has been used in inflammatory bowel disease CTs for over a decade.[16–23] It is important for CTs to stratify for erosive esophagitis severity using an accurate, reproducible grading system, and to have a quality control mechanism to ensure accurate application of the grading system both before and after any treatment. Adjudication processes should be robust and must address interobserver variability between adjudicators. We conducted a *post hoc* analysis of data from a Phase 2 randomized controlled trial of a new acid suppression therapy (AST) to evaluate the interobserver agreement between CT endoscopists and independent expert adjudicators from the International Working Group for the Classification of Oesophagitis (IWGCO).[24]

METHODS

Data collection

In a randomized, controlled CT (Trial Registration ClinicalTrials.gov: NCT05055128, EudraCT 2020–003319-91) of AST for ERE, three different doses of linaprazan glurate, a potassium channel acid blocker (PCAB), were compared with a PPI, lansoprazole, to document endoscopic healing of erosive esophagitis after four weeks of treatment.[24] CT endoscopists from 34 sites in eight countries reported the presence and severity of ERE, using the LA classification, before and after a four-week course of AST. CT endoscopists followed a central protocol that required using a high-quality medical video processor to record videos and images of the distal esophagus and gastroesophageal junction for all study participants, and required that videos include adequate coverage and time examining each area of interest to allow for LA grade assessment.

The endoscopic videos and images were uploaded to a web-based review platform, accessible by three central adjudicators, who were all blinded to patient information and treatment. Each video or image was initially reviewed independently by two adjudicators. Adjudicators could also pause the videos and review each frame as a still image. Adjudicators provided their assessment using a standard, online reporting form to report whether the video or image was of sufficiently good visual quality for evaluation and, if it was, to document the presence and severity of reflux esophagitis. A final diagnosis on LA grade was reached if both initial adjudicators documented the same LA grade for the video or image. In the event of disagreement, a third adjudicator independently reviewed the video or image; if the third reviewer's grade matched either of those reported by one of the initial two adjudicators, this was recorded as the final consensus grade. If a majority verdict was not reached after the first three reviews, all three adjudicators met by video conference to review the recording and document a final consensus diagnosis on the severity of esophagitis.

Statistical analysis: comparing CT endoscopists to adjudicators

Videos and images that were deemed by adjudicators to be of too poor quality for assessment were not included in the data analysis. For the remaining videos and images, the number of cases assessed as no esophagitis, LA grade A, B, C, or D by CT endoscopists and by adjudicators was cross-tabulated to determine the number of cases in which the CT endoscopist and adjudicators agreed or disagreed with respect to esophagitis severity. The extent of disagreement between CT endoscopists and adjudicators was calculated by assigning scores to each grade of esophagitis (e.g. no esophagitis, LA grade A, B, C, and D corresponded to scores of 0, 1, 2, 3, and 4, respectively), and subtracting the score from the CT endoscopist from the score of the adjudicators. For example, if the CT endoscopist's assessment was LA grade D esophagitis, and the adjudicators' assessment was no esophagitis, then the extent of disagreement would be −4.

Cohen's kappa (κ) was used to measure the strength of interobserver agreement between CT endoscopists and adjudicators on LA grade. Weighted kappa (κ_w) was used to account for the extent of disagreement when measuring the strength of interobserver agreement. Kappa values of <0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and >0.8 corresponded to poor, fair, moderate, good, and very good agreement, respectively. Univariable analysis was used to determine the association between prior PPI use, repeat procedure, country, or study site and interobserver agreement on LA grade; multivariable regression analysis was performed only for the effects of prior PPI use and repeat procedures due to the large number of countries and study sites.

Statistical analysis: comparing adjudicators to each other and to final diagnosis

For each adjudicator and for the final diagnosis, we recorded the number of cases assessed as no esophagitis, LA grade A, B, C or D. As with the analysis comparing CT endoscopists with adjudicators, the strength of interobserver agreement between adjudicators and with the final diagnosis was measured using Cohen's kappa, with weighted kappa used to account for the extent of disagreement with respect to the assigned LA grade. IBM Statistical Product and Service Solutions (SPSS) software Version 28.0.1.1 was used for all analyses in this study.

RESULTS

Analysis based on each individual LA grade of esophagitis

452 cases with videos or images from CT endoscopists were reviewed by adjudicators; of these, 64 cases had videos or images that were deemed too poor in quality for assessment. The remaining 388 cases were graded concordantly in 233 (60.1%), 133 (34.3%), and 22 cases (5.7%) after 2, 3, and 4 reviews, respectively.

Comparing CT endoscopists with adjudicators

Of the 388 cases with videos/images of sufficient quality for assessment, CT endoscopists graded 94 (24.2%), 117 (30.2%), 56 (14.4%), 102 (26.3%), and 19 (4.9%) cases as no esophagitis, LA grade A, B, C and D, respectively, whereas adjudicators graded 225 (58%), 43 (11.1%), 71 (18.3%), 31 (8.0%), and 18 (4.6%) cases as no esophagitis, LA grade A, B, C and D, respectively (Table 1 and Fig. 1).

CT endoscopists and adjudicators had agreement for 168 (43.3%) cases. CT endoscopists assigned a more severe grade than adjudicators in 185 (47.7%) cases, the extent of disagreement ranging by one (e.g. none to A, 115/29.6%), two (e.g. none to B, 26/6.7%), three (e.g. none to C, 41/10.6%), or four (e.g. none to D, 3/0.8%) grades (Fig. 2). There was fair agreement between CT endoscopists and adjudicators ($\kappa = 0.27$, 95% CI 0.21–0.32; $\kappa_w = 0.40$, 95% CI 0.34–0.46).

On univariable analysis, site ($P < 0.001$), country ($P < 0.001$), and repeat procedure ($P = 0.002$), but not prior PPI use ($P = 0.926$) was associated with higher interobserver agreement on LA grade. On multivariable analysis, repeat procedure (OR = 1.888, 95% CI 1.257–2.838, $P = 0.002$), but not prior PPI use (OR = 0.992, 95% CI 0.626–1.573, $P = 0.974$), was associated with higher interobserver agreement for LA grade.

Comparing adjudicators with each other and with final diagnosis

The strength of agreement was moderate between adjudicators A versus B ($\kappa = 0.43$, 95% CI 0.34–0.51), moderate between adjudicators A versus C ($\kappa = 0.47$, 95% CI 0.39–0.56), and moderate between adjudicators B versus C ($\kappa = 0.47$, 95% CI 0.38–0.55). The strength of agreement between the final diagnosis and each adjudicator was good (adjudicator A, $\kappa = 0.78$, 95% CI 0.72–0.84; adjudicator B, $\kappa = 0.78$, 95% CI 0.72–0.85; adjudicator C, $\kappa = 0.75$, 95% CI 0.69–0.81).

When accounting for the extent of disagreements, the strength of agreement was good between adjudicators A versus B ($\kappa_w = 0.66$, 95% CI 0.59–0.74), good between adjudicators A versus C ($\kappa_w = 0.63$, 95% CI 0.56–0.70), and good between adjudicators B versus C ($\kappa_w = 0.64$, 95% CI 0.57–0.71). The strength of agreement between the final diagnosis and each adjudicator was very good (adjudicator A, $\kappa_w = 0.87$, 95% CI 0.83–0.91; adjudicator B, $\kappa_w = 0.87$, 95% CI 0.83–0.91; adjudicator C, $\kappa_w = 0.84$, 95% CI 0.80–0.88).

DISCUSSION

To our knowledge, our study is the largest study to investigate interobserver agreement between CT endoscopists and expert adjudicators on assessing the LA grade of ERE based on endoscopic videos. CT endoscopists typically reported more severe LA grades compared to central adjudicators for the same videos/images, with the difference being one grade more severe for these cases. The exact reason for this finding is unclear. It is reasonable to assume that the adjudicators, compared to CT endoscopists, would have more experience using the LA classification and this experience would be expected to allow for a more accurate application of the grading system. It is unclear if LA assessments were influenced by the presence of minimal changes of esophagitis, such as erythema or friability. Minimal changes may have led to mistakenly increased grading of esophagitis

Table 1 Breakdown of disagreements between CT endoscopists and expert adjudicators

		Expert adjudicators					
		Normal	LA-A	LA-B	LA-C	LA-D	Total
CT endoscopists	Normal	87	4	3	0	0	94
	LA-A	74	28	14	1	0	117
	LA-B	22	7	21	4	2	56
	LA-C	39	2	31	23	7	102
	LA-D	3	2	2	3	9	19
	Total	225	43	71	31	18	388

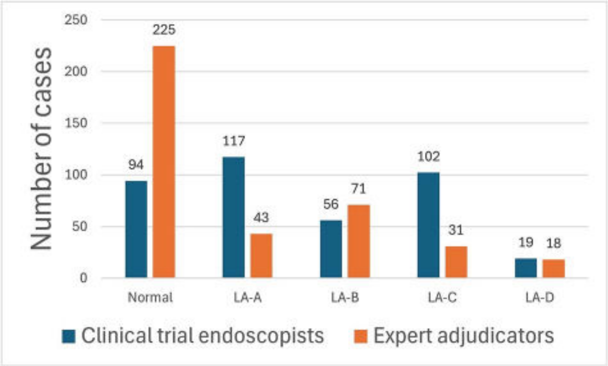

Fig. 1 Number of videos graded by CT endoscopists and adjudicators as no esophagitis, LA grade A, B, C, and D.

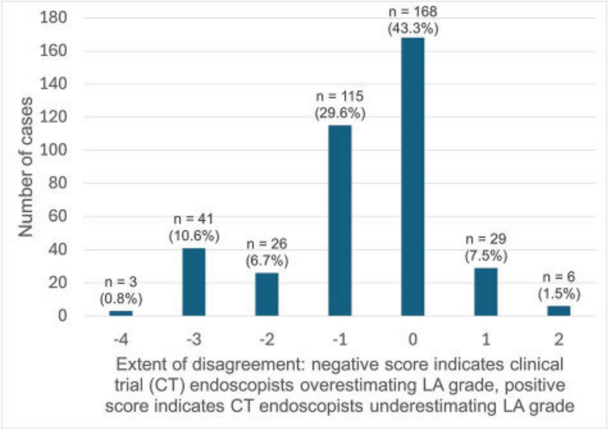

Fig. 2 Extent of disagreement between CT endoscopists and adjudicators on LA grade.

in this CT. However, previous studies have, also, shown poor interobserver agreement for minimal changes.[7,25–28]

There was only a fair agreement between CT endoscopists and adjudicators on LA grades, with $\kappa = 0.27$. The strength of agreement was higher for the weighted kappa ($\kappa_{\mathrm{W}} = 0.40$) which accounts for the extent of disagreement; however, the strength of agreement was still only fair. It is possible that the difference in endoscopic experience between the CT endoscopists and the adjudicators influenced the interobserver agreement. One prior study, in which endoscopists assessed still images, showed fair interobserver agreement for LA assessment with $\kappa = 0.26$ when comparing endoscopists who had performed more than 3000 procedures and less experienced endoscopists who had performed only 100–500 procedures.[29] Interobserver agreement on LA classification from most previous studies ranged from $\kappa = 0.22$ to $\kappa = 0.65$.[4,7,11,12,29–33] However, there is one study in which six attending physicians and three trainees assigned LA grades to still images; the interobserver agreement was lower at $\kappa = 0.22$, with $\kappa = 0.20$ for attending physicians and $\kappa = 0.31$ for trainees.[32]

The regression analysis shows that repeat procedure was associated with interobserver agreement. It is possible that the CT endoscopists benefited from a mild training effect by gaining experience in assessing LA grades on initial procedures, which led to improved ability using the LA classification for repeat procedures and thus improving interobserver agreement.[34] Conversely, prior PPI use was not associated with interobserver agreement on the LA classification. This may suggest that prior PPI use will not interfere with the reproducibility of LA assessments. However, the PPI dose, duration, and interval between last PPI use and enrolment in the study is unknown for the patients in this CT. It is unclear why a repeat procedure was associated with increased interobserver agreement on univariable but not multivariable regression analysis.

Comparing adjudicators with each other and with final diagnosis

The interobserver agreement between individual IWGCO expert adjudicators on LA assessments was only fair when including images/videos of quality too poor to assess, with κ ranging from 0.34 to 0.36. When we compared each individual adjudicator to the final diagnosis, the strength of interobserver agreement was improved from fair to good, with κ ranging from 0.69 to 0.77. This finding echoes a previous study in which three consultant endoscopists each individually assigned LA grades to 35 endoscopy videos, and then met together to discuss and reach a final consensus on the LA grades of each case.[33] Interobserver agreement between individual endoscopists was moderate with $\kappa = 0.58$, but when comparing endoscopists to the consensus decision, agreement improved to good with $\kappa = 0.77$.[33] Our study differs by having a much greater number of cases reviewed, with cases collected in the context of a CT. It is likely that the reason for stronger agreement when comparing endoscopists to the final decision as a consensus decision mitigates individual biases in using the LA assessment for grading esophagitis. These data suggest that central reading protocols for CTs should include a formal adjudication process to resolve disagreements between the initial reviewers, similar to the adjudication processes employed for evidence-based systematic reviews.[35]

While central adjudication of videos and images has been increasingly used in CTs involving inflammatory bowel disease, our study is one of few that investigate the interobserver agreement between CT endoscopists and central expert adjudicators on grading erosive esophagitis severity using the LA classification. Our study is similar to Spechler *et al.*, who found that there is only weak to moderate agreement between central adjudicators and CT endoscopists.[15] Our study differs in that almost all of the cases were videos, unlike the study reported by Spechler *et al.* which used only still images. Our study also compares interobserver agreement between experts, which is important as previous studies have shown interobserver differences between experts.[33] Central adjudication of clinical images has been shown to increase effect size and decrease placebo rates in CTs involving inflammatory bowel disease, and our results strongly support the use of central adjudication in CTs on endoscopic assessments of ERE.[16] It is noteworthy that the esophagitis severity reported by central expert adjudicators was frequently less than that reported by CT endoscopists, suggesting that there may have been a tendency, subconscious or otherwise, to 'upgrade' disease severity such that patients met study enrolment criteria.

Some of the strengths of this study are that the adjudicators were blinded to the patient, treatment, and the CT endoscopists' grading. There was a defined process for resolving discrepancies in adjudication. All cases were evaluated using video recordings in addition to still images to address concerns that the latter may not have demonstrated the entire region of interest. Our study has several limitations. There was no standard recording protocol for the videos and images, and the quality of the videos and images provided to the adjudicators varied widely. The videos submitted for evaluation may not have been representative of the views available to CT endoscopists during their live assessment of each patient. The experience and comfort level of CT endoscopists in using the LA classification is unknown. As this was a *post hoc* analysis of CT data, the adjudication process was not integrated into the CT from the beginning.

Future CTs on erosive esophagitis should be designed to ensure the accurate grading of esophagitis severity using the LA classification. Inaccurate classification of esophagitis severity can lead to inappropriate enrolment or treatment of patients and can lead to greater costs and greater risks of inconclusive studies. Several factors need to be considered, such as ensuring that videos or images are recorded with sufficiently high quality. CT endoscopists and central expert adjudicators should receive training on using the LA classification prior to and throughout the study, as such training can improve interobserver agreement.[22,36] A protocol should exist to mediate interobserver discrepancies between expert adjudicators. There should be a feedback mechanism to improve the assessments of CT endoscopists and expert adjudicators. These recommendations for evaluating the severity of reflux esophagitis should match, for example, the recommendations for CTs on inflammatory bowel disease.[16–23]

CONCLUSION

There was only a fair agreement between CT endoscopists and expert central adjudicators in grading reflux esophagitis severity using the LA classification. Further education in using the LA classification is needed to improve interobserver agreement. CTs in which endoscopy is used to determine subject eligibility or treatment response should directly incorporate adjudication processes into their design, preferably including independent, central reviewers and a conflict resolution, adjudication strategy.

ACKNOWLEDGMENT(S)

We appreciate the unrestricted funding from Cinclus Pharma to analyze and report this data.

References

1. Dent J, Becher A, Sung J, Zou D, Agréus L, Bazzoli F. Systematic review: patterns of reflux-induced symptoms and esophageal endoscopic findings in large-scale surveys. Clin Gastroenterol Hepatol 2012; 10: 863–873.e3. https://doi.org/10.1016/j.cgh.2012.02.028.
2. Ronkainen J, Talley N J, Storskrubb T *et al.* Erosive esophagitis is a risk factor for Barrett's esophagus: a community-based endoscopic follow-up study. Am J Gastroenterol 2011; 106: 1946–52. https://doi.org/10.1038/ajg.2011.326.
3. Lassen A, Hallas J, de Muckadell O B. Incidence and risk of esophageal adenocarcinoma - a population-based cohort study. Am J Gastroenterol 2006; 101: 1193–9. https://doi.org/10.1111/j.1572-0241.2006.00550.x.
4. Armstrong D, Bennett J R, Blum A L *et al.* The endoscopic assessment of esophagitis: a progress report on observer agreement. Gastroenterology. 1996; 111: 85–92. https://doi.org/10.1053/gast.1996.v111.pm8698230.
5. Katz P O, Dunbar K B, Schnoll-Sussman F H, Greer K B, Yadlapati R, Spechler S J. ACG clinical guideline for the diagnosis and management of gastroesophageal reflux disease. Am J Gastroenterol 2022; 117: 27–56. https://doi.org/10.14309/ajg.0000000000001538.
6. Iwakiri K, Fujiwara Y, Manabe N *et al.* Evidence-based clinical practice guidelines for gastroesophageal reflux disease 2021. J Gastroenterol 2022; 57: 267–85. https://doi.org/10.1007/s00535-022-01861-z.
7. Lundell L R, Dent J, Bennett J R *et al.* Endoscopic assessment of oesophagitis: clinical and functional correlates and further validation of the Los Angeles classification. Gut. 1999; 45: 172–80. https://doi.org/10.1136/gut.45.2.172.
8. Johnsson F, Weywadt L, Solhaug J H, Hernqvist H, Bengtsson L. One-week omeprazole treatment in the diagnosis of gastro-oesophageal reflux disease. Scand J Gastroenterol 1998; 33: 15–20. https://doi.org/10.1080/00365529850166149.
9. Adachi K, Fujishiro H, Katsube T *et al.* Predominant nocturnal acid reflux in patients with Los Angeles grade C and D reflux esophagitis. J Gastroenterol Hepatol 2001; 16: 1191–6. https://doi.org/10.1046/j.1440-1746.2001.02617.x.
10. Yadlapati R, Hubscher E, Pelletier C *et al.* Induction and maintenance of healing in erosive esophagitis in the United States. Expert Rev Gastroenterol Hepatol 2022; 16: 967–80. https://doi.org/10.1080/17474124.2022.2134115.
11. Rath H C, Timmer A, Kunkel C *et al.* Comparison of interobserver agreement for different scoring systems for reflux esophagitis: impact of level of experience. Gastrointest Endosc 2004; 60: 44–9. https://doi.org/10.1016/s0016-5107(04)01289-1.
12. Pandolfino J E, Vakil N B, Kahrilas P J. Comparison of inter- and intraobserver consistency for grading of esophagitis by expert and trainee endoscopists. Gastrointest Endosc 2002; 56: 639–43. https://doi.org/10.1067/mge.2002.129220.
13. Gustavsson S, Bergström R, Erwall C, Krog M, Lindholm C E, Nyrén O. Reflux esophagitis: assessment of therapy effects and observer variation by video documentation of endoscopy findings. Scand J Gastroenterol 1987; 22: 585–91. https://doi.org/10.3109/00365528708991902.
14. Bytzer P, Havelund T, Hansen J M. Interobserver variation in the endoscopic diagnosis of reflux esophagitis. Scand J Gastroenterol 1993; 28: 119–25. https://doi.org/10.3109/00365529309096057.
15. Spechler S J, Laine L, DeVault K R, Nabulsi A, Hunt B, Katz P. Comparison of Los Angeles grades of erosive esophagitis scored by local investigators vs central adjudicators in a clinical trial. Clin Gastroenterol Hepatol 2024; 22: 2526–2528.e1. https://doi.org/10.1016/j.cgh.2024.05.007.
16. Gottlieb K, Daperno M, Usiskin K *et al.* Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments. Gut. 2021; 70: 418–26. https://doi.org/10.1136/gutjnl-2020-320690.
17. Gottlieb K, Travis S, Feagan B, Hussain F, Sandborn W J, Rutgeerts P. Central reading of endoscopy endpoints in inflammatory bowel disease trials. Inflamm Bowel Dis 2015; 21: 2475–82. https://doi.org/10.1097/mib.0000000000000470.
18. Khanna R, Ma C, Jairath V, Vande Casteele N, Zou G, Feagan B G. Endoscopic assessment of inflammatory bowel disease activity in clinical trials. Clin Gastroenterol Hepatol 2022; 20: 727–736.e2. https://doi.org/10.1016/j.cgh.2020.12.017.
19. Hanzel J, Jairath V, De Cruz P *et al.* Recommendations for standardizing clinical trial design and endoscopic assessment in postoperative Crohn's disease. Inflamm Bowel Dis 2022; 28: 1321–31. https://doi.org/10.1093/ibd/izab259.
20. Raine T, Pavey H, Qian W *et al.* Establishment of a validated central reading system for ileocolonoscopy in an academic setting. Gut. 2022; 71: 661–4. https://doi.org/10.1136/gutjnl-2021-325575.
21. Feagan B G, Sandborn W J, D'Haens G *et al.* The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. Gastroenterology. 2013; 145: 149–157.e2. https://doi.org/10.1053/j.gastro.2013.03.025.
22. Daperno M, Comberlato M, Bossa F *et al.* Training programs on endoscopic scoring systems for inflammatory bowel disease lead to a significant increase in interobserver agreement among community gastroenterologists. J Crohns Colitis 2017; 11: 556–61. https://doi.org/10.1093/ecco-jcc/jjw181.
23. Buchner A M, Farraye F A, Iacucci M. AGA clinical practice update on endoscopic scoring systems in inflammatory bowel disease: commentary. Clin Gastroenterol Hepatol 2024; 22: 2188–96. https://doi.org/10.1016/j.cgh.2024.06.048.
24. Sharma P, Vaezi M, Unge P, Andersson K, Larsson K, Popadiyn I, Rosenholm M, Rosztoczy A, Yektaei E, Armstrong D. A dose-finding study of linaprazan glurate, a novel potassium-competitive acid blocker, vs. lansoprazole for the treatment of erosive esophagitis – a randomized clinical trial. Aliment Pharmacol Ther 2025;61(10):1590–1602. https://doi.org/10.1111/apt.70109.
25. Lee S P, Kae S H, Jang H J, Koh D H, Jung E S. Inter-observer variability of experts and trainees for the diagnosis of reflux esophagitis: comparison of linked color imaging, blue laser imaging, and white light imaging. J Dig Dis 2021; 22: 425–32. https://doi.org/10.1111/1751-2980.13023.
26. Miwa H, Yokoyama T, Hori K *et al.* Interobserver agreement in endoscopic evaluation of reflux esophagitis using a modified Los Angeles classification incorporating grades N and M: a validation study in a cohort of Japanese endoscopists. Dis Esophagus 2008; 21: 355–63. https://doi.org/10.1111/j.1442-2050.2007.00788.x.
27. Amano Y, Ishimura N, Furuta K *et al.* Interobserver agreement on classifying endoscopic diagnoses of nonerosive esophagitis. Endoscopy. 2006; 38: 1032–5.
28. Edebo A, Tam W, Bruno M *et al.* Magnification endoscopy for diagnosis of nonerosive reflux disease: a proposal of diagnostic criteria and critical analysis of observer variability. Endoscopy. 2007; 39: 195–201.
29. Kusano M, Ino K, Yamada T *et al.* Interobserver and intraobserver variation in endoscopic assessment of GERD using the "Los Angeles" classification. Gastrointest Endosc 1999; 49: 700–4. https://doi.org/10.1016/s0016-5107(99)70285-3.
30. Lee Y C, Lin J T, Chiu H M *et al.* Intraobserver and interobserver consistency for grading esophagitis with narrow-band imaging. Gastrointest Endosc 2007; 66: 230–6. https://doi.org/10.1016/j.gie.2006.10.056.
31. Lee S H, Jang B I, Kim K O *et al.* Endoscopic experience improves interobserver agreement in the grading of esophagitis by Los Angeles classification: conventional endoscopy and optimal band image system. Gut Liver 2014; 8: 154–9. https://doi.org/10.5009/gnl.2014.8.2.154.
32. Nasseri-Moghaddam S, Razjouyan H, Nouraei M *et al.* Inter- and intra-observer variability of the Los Angeles classification: a reassessment. Arch Iran Med 2007; 10: 48–53.
33. Wong R K, Yeoh K G, Gwee K A, Tay H W, Ho K Y. Validation of structured scoring using the LA classification for

esophagitis and endoscopically suspected Barrett's esophagus in a tertiary Asian endoscopy center. Validation study. Journal of Gastroenterology & Hepatology 2009; 24: 103–6. https://doi.org/10.1111/j.1440-1746.2008.05680.x.

34. Jin E H, Chung S J, Lim J H *et al.* Training effect on the inter-observer agreement in endoscopic diagnosis and grading of atrophic gastritis according to level of endoscopic experience. J Korean Med Sci 2018; 33: e117. https://doi.org/10.3346/jkms.2018.33.e117.

35. Polanin J R, Pigott T D, Espelage D L, Grotpeter J K. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods 2019; 10: 330–42. https://doi.org/10.1002/jrsm.1354.

36. Dadgar K, Wang D, Yuan Y, Sinclair P, Sharma P, Armstrong D. Training and assessment of Clinician's utilization of the Los Angeles classification for reflux esophagitis. Foregut. 2025; 5: 27–34 . https://doi.org/10.1177/26345161241269002.