ORIGINAL ARTICLE

Reliability and responsiveness of endoscopic disease activity assessment in eosinophilic esophagitis

Christopher Ma, MD, MPH,[1,2] Albert J. Bredenoord, MD, PhD,[3] Evan S. Dellon, MD, MPH,[4] Jeffrey A. Alexander, MD,[5] Luc Biedermann, MD,[6] Malcolm Hogan, MSc,[2] Leonardo Guizzetti, PhD,[2] Guangyong Zou, PhD,[2,7] David A. Katzka, MD,[5] Mirna Chehade, MD, MPH,[8] Gary W. Falk, MD, MS,[9] Glenn T. Furuta, MD,[10] Sandeep K. Gupta, MD,[11] Amir F. Kagalwalla, MD,[12,13] Alain M. Schoepfer, MD,[14] Stephan Miehlke, MD, PhD,[15] Fouad J. Moawad, MD,[16] Kathryn Peterson, MD, MS,[17] Nirmala P. Gonsalves, MD,[18] Alex Straumann, MD,[19] Joshua B. Wechsler, MD, MS,[12] Julie Rémillard, MSc,[2] Lisa M. Shackelton, PhD,[2] Hector S. Almonte, MD,[18] Brian G. Feagan, MD,[2,7,20] Vipul Jairath, MD, PhD,[2,7,20] Ikuo Hirano, MD[18]

Calgary, Alberta; London, Ontario, Canada; Amsterdam, The Netherlands; Chapel Hill, North Carolina; Rochester, Minnesota; New York, New York; Philadelphia, Pennsylvania; Aurora, Colorado; Indianapolis, Indiana; Chicago, Illinois; La Jolla, California; Salt Lake City, Utah, USA; Zurich, Lausanne, Switzerland; Hamburg, Germany

Background and Aims: Endoscopic outcomes have become important measures of eosinophilic esophagitis (EoE) disease activity, including as an endpoint in randomized controlled trials (RCTs). We evaluated the operating properties of endoscopic measures for use in EoE RCTs.

Methods: Modified Research and Development/University of California Los Angeles appropriateness methods and a panel of 15 international EoE experts identified endoscopic items and definitions with face validity that were used in a 2-round voting process to define simplified (all items graded as absent or present) and expanded versions (additional grades for edema, furrows, and/or exudates) of the EoE Endoscopic Reference Score (EREFS). Inter- and intrarater reliability of these instruments (expressed as intraclass correlation coefficients [ICC]) were evaluated using paired endoscopy video assessments of 2 blinded central readers in patients before and after 8 weeks of proton pump inhibitors, swallowed topical corticosteroids, or dietary elimination. Responsiveness was measured using the standardized effect size (SES).

Results: The appropriateness of 41 statements relevant to EoE endoscopic activity (endoscopic items, item definitions and grading, and other considerations relevant for endoscopy) was considered. The original and expanded EREFS demonstrated moderate-to-substantial inter-rater reliability (ICCs of .472-.736 and .469-.763, respectively) and moderate-to-almost perfect intrarater reliability (ICCs of .580-.828 and .581-.828, respectively). Strictures were least reliably assessed (ICC, .072-.385). The original EREFS was highly responsive (SES, 1.126 [95% confidence interval {CI}, .757-1.534]), although both expanded versions of EREFS, scored based on worst affected area, were numerically most responsive to treatment (expanded furrows: SES, 1.229 [95% CI, .858-1.643]; all items expanded: SES, 1.252 [95% CI, .880-1.667]). The EREFS and its modifications were not more reliably scored by segment and also not more responsive when proximal and distal EREFSs were summed.

Conclusions: EREFS and its modifications were reliable and responsive, and the original or expanded versions of the EREFS may be preferred in RCTs. Disease activity scored based on the worst affected area optimizes reliability and responsiveness. (Gastrointest Endosc 2022;■:1-12.)

(footnotes appear on last page of article)

Swallowed topical corticosteroids and proton pump inhibitors remain the mainstay of medical therapy for eosinophilic esophagitis (EoE), although multiple novel pharmacologic therapies are currently under evaluation.[1] Historically, drug development for EoE has been hampered by uncertainty regarding endpoints for measuring disease activity and response to treatment, resulting in a lack of standardized outcome measures for randomized controlled trials (RCTs) that support labeling claims. Regulatory guidance currently requires

achievement of co-primary endpoints that consider both patient-reported outcomes and histologic remission for new therapies in development for EoE.[2] The role for endoscopy in EoE RCTs is less clear. Endoscopy is required for obtaining biopsy specimens for histologic assessment, is an objective method for evaluation of disease activity, and has been recognized as a critical component for inclusion in a recent EoE core outcome set.[3] However, validation of endoscopic instruments for use in EoE RCTs is required.

Early research in this field used nonvalidated global assessments of endoscopic appearance based on common endoscopic findings such as esophageal rings, linear furrowing, mucosal pallor/edema, and white plaques,[4] which formed the basis for the development and validation of a novel endoscopic classification and grading instrument described by Hirano et al[5] (the EoE Endoscopic Reference Score [EREFS]; Supplementary Table 1, available online at www.giejournal.org). Although secondary endpoints based on the EREFS have been incorporated in EoE trials,[3,6] several challenges exist for the use of endoscopy in RCTs. First, consensus is lacking on the most appropriate methods for scoring the EREFS, including whether different segments of the esophagus should be scored separately, whether item grading can be improved, and whether component items should be summed or weighted.[7,8] Second, evaluation of endoscopic activity in RCTs by the local site endoscopist introduces potential observation bias because of lack of blinding to patient symptoms and trial time point, although the reliability of central assessment of endoscopic features of EoE has not been well established. Finally, identification of the most responsive endoscopic features of EoE is needed to efficiently detect treatment effects compared with placebo. Therefore, we undertook a multiple-phase study to systematically evaluate the performance of endoscopic evaluation for use in EoE RCTs by defining appropriate instruments and items with face validity for endoscopic assessment, evaluating the inter- and intrarater reliability of blinded central assessment of endoscopic activity, and measuring the responsiveness to change and longitudinal validity of EoE endoscopic features after treatment.

METHODS

Overall study design

First, a panel of international EoE experts was assembled, and modified Research and Development University of California Los Angeles appropriateness methodology (RAM)[9] was used to assess the face validity and feasibility of different approaches to endoscopic assessment in EoE. Next, endoscopic videos from patients with EoE before and after treatment with swallowed topical corticosteroids, proton pump inhibitor, or elimination diet were collected and assessed by 2 independent blinded central readers in duplicate and random order using appropriate items and methods identified by the RAM panel. The reliability and responsiveness of endoscopic features of EoE were estimated based on these assessments (Fig. 1).

Modified RAM

The RAM combines best available evidence and expert experience in iterative rounds of voting and discussion. An international panel of 15 EoE experts was selected (10 adult and 5 pediatric gastroenterologists). The panelists have a diverse range of clinical and research practices, although all have extensive expertise in EoE and its endoscopic assessment. All panelists have published >30 peer-reviewed articles on EoE and were specifically selected for their content expertise. After an initial teleconference, a list of 41 statements informed by a systematic literature review[10] was circulated and included assessment of existing endoscopic indices and items, item definitions, item-level grading, location for disease evaluation, and other relevant considerations. Panelists anonymously rated the appropriateness of each statement on a 9-point Likert scale. Each statement was categorized as appropriate, uncertain, or inappropriate based on 2 components: the median panel rating and degree of disagreement (defined as having $\geq$5 panelists in both the lowest [1-3] and highest [7-9] 3-point regions). Results from the first-round survey were reviewed in a second moderated videoconference to discuss rationale for responses and clarify areas of disagreement. A revised survey was then recirculated for final voting.

In contrast to a Delphi process, the RAM does not force a consensus. This is important for initial steps in index development and validation because it allows the panel to explore and consider a wide variety of items and allows consideration of items of "uncertain" appropriateness to be further evaluated. A Delphi process is not desirable for these purposes because forcing a consensus may exclude potentially appropriate items that have not yet been rigorously tested in the literature. Rather, the purpose of the RAM is to include a panel of experts who have extensive experience in EoE to generate valid opinions regarding the potential value of different items, with any individual biases counteracted by discussing points of contention within the larger group yet preserving anonymized voting. The RAM process has been used extensively in the literature as a robust method for assessing appropriateness.

Endoscopic material

Clinical data and upper endoscopy videos were obtained from 2 prospective clinical studies conducted at Northwestern University and Amsterdam Medical Center. Adult ($\geq$18 years) EoE patients treated with swallowed topical budesonide/fluticasone (n = 16), proton pump inhibitors (n = 6), or elimination diet (n = 18) with video-

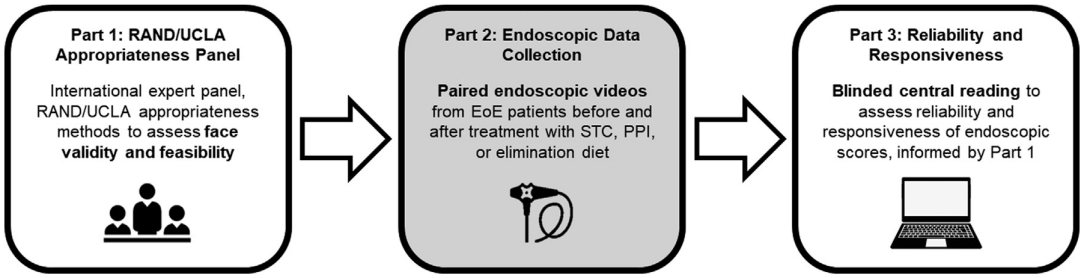


Figure 1. Study design. *RAND/UCLA*, Research and Development University of California Los Angeles; *EoE*, eosinophilic esophagitis; *STC*, swallowed topical corticosteroids; *PPI*, proton pump inhibitor.

recorded endoscopy performed at baseline and after 8 weeks of therapy were included. All videos included both insertion and withdrawal phases. Endoscopic dilation was masked to avoid confounding interpretation of impassable strictures or rings. All videos were evaluated by 2 expert EoE gastroenterologists (J.A.A. and L.B.) blinded to all clinical information. Videos were rescored at least 2 weeks from the first assessment in a different random order to facilitate memory extinction.

Endoscopic measures

The EREFS and all appropriate and uncertain items identified in the RAM process were used to evaluate endoscopic features of EoE, including the original EREFS (a 10-point scale including edema [0-2], exudates [0-2], furrows [0-2], rings [0-3], and stricture [absent or present]) and modifications based on collapsed or expanded definitions (see Results). Videos were rated independently for each item and separately scored based on the worst disease location and by esophageal segment (proximal [upper half], distal [lower half], and gastroesophageal junction [0-3 cm proximal to squamocolumnar junction]). Finally, a global measure of disease activity was assessed using a 100-mm visual analog scale (VAS), ranging from 0 (no endoscopic disease activity) to 100 (worst endoscopic disease activity ever seen). The VAS serves as an external anchor and is an accepted method for clinical outcome assessment at the regulatory level.[11] An inflammatory (sum of exudate, edema, and furrows) and fibrostenotic (sum of rings and strictures) subscore was calculated post-hoc using established definitions in the literature.[7,8]

Statistical methods

Reliability of endoscopic assessment. Both interrater (reliability between readers) and intrarater (reliability within a reader) reliability was evaluated in the agreement component of this study. Reliability was quantified by the intraclass correlation coefficient (ICC) for agreement rather than anchored against an external "criterion standard" used for diagnostic studies. Point estimates were derived using a 2-way random-effects analysis of variance model at baseline and follow-up. The 95% confidence interval (CI) was obtained using nonparametric cluster bootstrapping. Reliability estimates were interpreted according to benchmarks proposed by Landis and Koch[12] (poor ICC, <.00; slight, .00-.20; fair, .21-.40; moderate, .41-.60; substantial, .61-.80; or almost perfect, .81-1.00). The ICC is equivalent to the weighted κ for ordinal categorical data.[13] The ICC is more appropriate to use in this study because it gives more weight to extreme disagreements that are likely to be clinically relevant compared with the unweighted κ.[14]

Responsiveness and longitudinal validity. Responsiveness was evaluated as the ability to detect change after treatment. In the absence of a criterion standard definition of endoscopic improvement in EoE, we used a criterion for change defined by an improvement of at least one-half standard deviation (SD) in the overall baseline disease VAS. This is a validated threshold used for change discrimination in chronic diseases, has been accepted by regulators,[11] and has been previously used as the standard for instrument development for other gastroenterological conditions.[15-19] Standardized effect size (SES) and 2-sided 95% CIs were calculated. The SES is calculated by the mean difference between changed and unchanged groups divided by the SD of pooled observations; an SES of 0 indicates no difference between groups, whereas values >0 indicate larger group differences. SES estimates were interpreted according to Cohen's benchmarks (.2, small; .5, moderate; and .8, large effect size).[20] Responsiveness was also quantified nonparametrically using the probability for distinguishing patients with improvement from those without improvement, expressed as the area under the receiver-operating characteristic curve (AUC). The null value of differences between groups for the AUC is .5.

Longitudinal validity was evaluated using weighted correlation coefficients for changes in the EREFS and its modifications with changes in the VAS. We also evaluated the correlation between changes in EREFS and its modifications with changes in peak eosinophil count, recognizing that endoscopic activity is distinct from histologic inflammation,[21] and for the purposes of RCTs, patient-reported, endoscopic, histologic, and quality of life outcomes should all be measured.[22]

TABLE 1. Voting summary from the modified Research and Development University of California Los Angeles appropriateness methodology panel

Item	Median panel rating (interquartile range)	Rating 1-3 n (%)	Rating 4-6 n (%)	Rating 7-9 n (%)	Appropriateness
The preferred term "rings" is sufficient to identify this endoscopic item.	9 (8-9)	0 (.0)	0 (.0)	15 (100.0)	Appropriate
The preferred term "exudates" is sufficient to identify this endoscopic item.	9 (8-9)	0 (.0)	1 (7.7)	12 (92.3)	Appropriate
The preferred term "furrows" is sufficient to identify this endoscopic item.	8 (7-9)	1 (6.7)	0 (.0)	14 (93.3)	Appropriate
The preferred term "edema" is sufficient to identify this endoscopic item.	9 (8-9)	1 (6.7)	2 (13.3)	12 (80.0)	Appropriate
The preferred term "stricture" is sufficient to identify this endoscopic item.	8 (7-9)	2 (15.4)	1 (7.7)	10 (76.9)	Appropriate
Crepe paper esophagus should not be included as a component of an endoscopic instrument.	8 (7-9)	0 (.0)	1 (8.3)	11 (91.7)	Appropriate
Narrow-caliber esophagus should not be included as a component of an endoscopic instrument.	8 (7-9)	3 (27.3)	1 (9.1)	7 (63.6)	Appropriate
Ankylosaurus back sign should not be included as a component of an endoscopic instrument.	7 (3-8)	0 (.0)	0 (.0)	12 (100.0)	Appropriate
Pull sign should not be included as a component of an endoscopic instrument.	8.5 (8-9)	1 (7.1)	2 (14.3)	11 (78.6)	Appropriate
Stricture length should be assessed.	8 (7-9)	0 (.0)	3 (20.0)	12 (80.0)	Appropriate
Stricture length should be assessed as <1 cm (focal) or >1 cm.	9 (8-9)	5 (33.3)	4 (26.7)	6 (40.0)	Uncertain
Estimation of stricture diameter should be based on the size of the endoscope used in the procedure.	8 (7-9)	0 (.0)	0 (.0)	15 (100.0)	Appropriate
To facilitate the estimation of stricture diameter, the diameter of the endoscope should be recorded by the endoscopist.	6 (2-8)	0 (.0)	0 (.0)	15 (100.0)	Appropriate
To facilitate the estimation of stricture diameter, the manufacturer make and model should be recorded by the endoscopist.	8 (8-9)	1 (6.7)	6 (40.0)	8 (53.3)	Appropriate
Estimation of stricture diameter should be based on the inability to pass an adult (<10 mm) or pediatric (<5 mm) endoscope.	8 (8-9)	0 (.0)	5 (33.3)	10 (66.7)	Appropriate
Estimation of stricture diameter should be made based on knowledge of endoscope size and within the following ranges: <5 mm, 5-10 mm, 11-15 mm, >15 mm.	7 (4-8)	1 (6.7)	8 (53.3)	6 (40.0)	Uncertain
Exudates should be assessed on endoscope insertion rather than withdrawal.	7 (6-8)	1 (7.7)	2 (15.4)	10 (76.9)	Appropriate
For the purpose of clinical trials, a color graphic atlas should be used to provide visual examples of each item.	6 (5-7)	0 (.0)	0 (.0)	15 (100.0)	Appropriate
Assessment of endoscopic features should be performed as a global evaluation based on the most severe findings.	8 (3-9)	4 (26.7)	1 (6.7)	10 (66.7)	Appropriate
Assessment of endoscopic features should be performed in the proximal and distal esophagus.	7 (7-9)	0 (.0)	3 (20.0)	12 (80.0)	Appropriate
The proximal esophagus is defined as the upper half of the esophagus.	8 (8-9)	1 (6.7)	1 (6.7)	13 (86.7)	Appropriate
The distal esophagus is defined as the lower half of the esophagus.	8 (8-9)	1 (6.7)	1 (6.7)	13 (86.7)	Appropriate
Assessment of endoscopic features should be performed in the proximal and distal esophagus and the gastroesophageal junction (0-3 cm proximal to the squamocolumnar junction).	7 (3-8)	4 (26.7)	3 (20.0)	8 (53.3)	Appropriate
Grading of exudates should be as previously described for the EREFS where 0 = none, 1 = mild (lesions involving <10% of the esophageal surface area), 2 = severe (lesions involving >10% of the esophageal surface area).	7 (4-9)	3 (20.0)	3 (20.0)	9 (60.0)	Appropriate
Grading of exudates should be as previously described for the EREFS but should also include a score of 3, where 0 = none, 1 = mild (lesions involving <10% of the esophageal surface area), 2 = moderate (lesions involving >10% and <25% of the esophageal surface area), 3 = severe (lesions involving >25% of the esophageal surface area).	5 (3-7)	5 (33.3)	6 (40.0)	4 (26.7)	Uncertain

(*continued on the next page*)

TABLE 1. Continued

Item	Median panel rating (interquartile range)	Rating 1-3 n (%)	Rating 4-6 n (%)	Rating 7-9 n (%)	Appropriateness
In preference to a scale of 0 to 1 (absent vs present), grading of furrows should be on a scale of 0 to 2, where 0 = absent, 1 = mild (vertical lines present without visible depth), 2 = severe (vertical lines with mucosal depth [indentation]).	8 (5-8)	3 (20.0)	2 (13.3)	10 (66.7)	Appropriate
In preference to a scale of 0 to 1 (where 0 = absent [distinct vascularity is present] and 1 = present [loss of clarity and absence of vascular markings]), scoring of edema should be on a scale of 0 to 2, where 0 = absent (distinct vascularity is present), 1 = reduced or loss of clarity of vascular markings, 2 = absence of vascular markings.	5 (3-8)	6 (40.0)	2 (13.3)	7 (46.7)	Uncertain

EREFS, Eosinophilic Esophagitis Endoscopic Reference Score.

Sample size and ethical considerations. The sample size required for this study was informed by the minimum acceptable standard for inter-rater reliability for use in an RCT setting. Assuming a true ICC of .75 and with analysis by a 1-way random-effects model, evaluation of 80 videos by 2 readers yielded >80% power for obtaining the 1-sided 95% lower bound for the ICC of >.60 (substantial reliability by Landis and Koch benchmarks).[23] All analyses were performed using SAS, version 9.4 (SAS Institute, Cary, NC, USA). Research ethics board approval was obtained from Northwestern University and the Amsterdam University Medical Center. All participants consented to secondary use of data, which was anonymized for use in this study. All readers consented to participate in the study.

RESULTS

RAM panel results

Endoscopic items. Appropriateness ratings from the RAM process are summarized in Table 1. Major components of the EREFS were considered appropriate for assessment in EoE RCTs, with the preferred terms (rings, exudates, furrows, edema, and stricture) sufficient for identifying the relevant findings rather than synonymous terms (eg, trachealization, corrugated esophagus, plaques/spots). The panel experts noted that some features are rare or incorporated in other items (eg, ankylosaurus back sign),[24] may be too subjective (eg, pull sign, crêpe-paper esophagus), or would be better evaluated with modalities other than endoscopy (eg, narrow-caliber esophagus). Although estimation of stricture length was considered appropriate, there was uncertainty whether this metric should be dichotomized (<1 cm or ≥1 cm), because it may be challenging to distinguish from an impassable concentric ring. Conversely, estimation of continuous stricture length may be infeasible, especially for central readers.

Item grading and esophageal segments. Expanded or condensed grading of the EREFS items was considered. Condensed grading may simplify assessment and potentially reduce interobserver variability. Expanded grading was believed to increase the dynamic range, with the potential to improve responsiveness for detection of subtle endoscopic changes after treatment. Grading of exudates should be done on insertion to avoid disruption by endoscope passage. The panel preferred grading furrows using a scale of 0 to 2 (0 = absent, 1 = mild [vertical lines without visible depth], 2 = severe [vertical lines with mucosal depth/indentation]) rather than dichotomizing to absent or present. There was disagreement whether edema should be dichotomized or scored on a scale of 0 to 2 (0 = distinct vascularity present, 1 = reduced/loss of vascular marking clarity, 2 = absence of vascular markings). Assessment of endoscopic features based on the most severe finding was considered appropriate, although separate segment scores (proximal/distal esophagus, gastroesophageal junction) might capture the extent of disease activity more accurately.

Scoring of the EREFS. Several versions of the EREFS were evaluated for reliability and responsiveness based on the results of the RAM study (Table 2). We reported results for the original EREFS (range, 0-8) without minor features, simplified EREFS (all items assessed as absent or present; range, 0-5), furrows-expanded EREFS (furrows graded as 0-2; range, 0-9), and fully expanded EREFS (exudates graded as 0-3, edema graded as 0-2, and furrows graded as 0-2; range, 0-11).

Patient demographics

Endoscopy videos from 41 patients with EoE were included (n = 82). Demographic characteristics are summarized in Supplementary Table 2 (available online at www.giejournal.org). The mean peak eosinophil count was 61.8 (SD, 23.3) eosinophils per high-power field at enrollment and 12.6 (SD, 20.4) post-treatment. The mean (original) EREFS score was 4.4 (SD, 1.6) at enrollment and 3.0 (SD, 1.7) post-treatment (Supplementary Fig. 1, available online at www.giejournal.org).

Reliability

The ICCs for the original EREFS ranged from .736 (95% CI, .518-837) at baseline to .602 (95% CI, .410-744) post-treatment when calculated based on assessment of the

TABLE 2. Adaptations of the EREFS classification considered

Feature	Original EREFS (range, 0-8)	Simplified EREFS (range, 0-5)	Expanded furrows EREFS (range, 0-9)	Fully expanded EREFS (range, 0-11)
Exudates	Grade 0: None Grade 1: Mild Grade 2: Severe	**Grade 0: Absent** **Grade 1: Present**	Grade 0: None Grade 1: Mild Grade 2: Severe	**Grade 0: None** **Grade 1: Mild** **Grade 2: Moderate** **Grade 3: Severe**
Rings	Grade 0: None Grade 1: Mild Grade 2: Moderate Grade 3: Severe	**Grade 0: Absent** **Grade 1: Present**	Grade 0: None Grade 1: Mild Grade 2: Moderate Grade 3: Severe	Grade 0: None Grade 1: Mild Grade 2: Moderate Grade 3: Severe
Edema	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present	**Grade 0: None** **Grade 1: Mild** **Grade 2: Severe**
Furrows	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present	**Grade 0: Absent** **Grade 1: Mild** **Grade 2: Severe**	**Grade 0: Absent** **Grade 1: Mild** **Grade 2: Severe**
Stricture	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present	Grade 0: Absent Grade 1: Present

Definitions of individual items are defined in Table 1; differences from the original EREFS are bold.
EREFS, Eosinophilic Esophagitis Endoscopic Reference Score.

worst finding for each feature, consistent with moderate-to-substantial inter-rater reliability (Table 3). Moderate-to-substantial inter-rater reliability was also observed for the simplified EREFS (ICC, .519-.694) and both expanded EREFS versions (ICC, .601-.763). The intrarater reliability for the original, simplified, and both expanded EREFS versions was moderate to almost perfect (intrarater ICC, .594-.828).

The reliability of individual endoscopic items is summarized in Table 4. Rings were most reliably assessed (moderate-to-substantial inter-rater ICC, .442-.688). There was fair-to-moderate reliability (inter-rater ICC, .212-.573) for exudates; the original (0-2) and simplified (0-1) grading versions were more reliably assessed than the expanded version (grading, 0-3). There was fair-to-moderate inter-rater reliability for evaluation of edema and furrows (inter-rater ICC, .312-.536). Strictures were the least reliably assessed feature: fair (ICC, .385 [95% CI, .090-628]) at baseline and slight (ICC, .072 [95% CI, .000-244]) post-treatment. Accordingly, inter-rater reliability for stricture diameter was also poor (ICC, .055-.187). All endoscopic items had at least moderate intrarater reliability regardless of grading method. The intrarater reliability for exudates and rings ranged from moderate to almost perfect. Similarly, both the inflammatory and fibrostenotic subscores were reliably assessed, with moderate-to-substantial inter-rater reliability (ICC, .481-.0.784) and substantial-to-almost perfect intrarater reliability (ICC, .696-.814).

Responsiveness

The original, simplified, and expanded versions of the EREFS, scored in the worst disease location, were highly responsive to treatment (SES, .974-1.252; AUC, .765-.816) (Table 5). The fully expanded EREFS version was generally most responsive when features were assessed in the worst affected area (SES, 1.252 [95% CI, .880-1.667]; AUC, .806 [95% CI-712-875]). Responsiveness was highest when using the global rating based on the worst affected area and generally higher in the distal esophagus and gastroesophageal junction than the proximal esophagus. When the EREFS was calculated as the sum of scores from the proximal and distal esophagus, it was numerically less responsive compared with scoring in the worst affected area (SES of .933 [95% CI, .566-1.329] vs 1.126 [95% CI, .757-1.534], respectively). Changes in all EREFS versions scored in the worst area were correlated with the change in VAS over time (weighted global correlation r = .796-.911). Changes in the instruments were only weakly correlated with changes in peak eosinophil count (Table 5).

Overall, exudates (SES-874-.968; AUC, .702-.723) and furrows (SES, .789-1.036; AUC, .686-.723) had the largest effect sizes (Table 6). In contrast, rings and strictures had mostly small effect sizes and weak correlations with the change in VAS ($r < .50$). Simplified (absent or present) grading of items was generally less responsive than the original grading system.

The inflammatory subscore (SES, 1.217 [95% CI, .845-1.631]; AUC, .787 [95% CI, .661-875]) was numerically but not significantly more responsive compared with the fibrostenotic subscore (SES, .558 [95% CI, .196-938]; AUC, .644 [95% CI, .480-779]; $P = .187$). A sensitivity analysis that excluded videos with endoscopic dilation (eg, with impassable strictures or rings requiring dilation) resulted in greater fibrostenotic subscore responsiveness (SES, .890 [95% CI, .269-1.654]; AUC, .723 [95% CI, .549-849]).

DISCUSSION

In this study, we aimed to improve the application of endoscopic assessment in EoE RCTs by assessing the

TABLE 3. Inter- and intrarater reliability of the EREFS and its modifications

Total EREFS score		Reliability intraclass correlation coefficient (95% confidence interval)			
		Baseline		Post-treatment	
Segment	EREFS version	Inter-rater	Intrarater	Inter-rater	Intrarater
Worst disease (global)	Original	.736 (.518-.837)	.828 (.656-.912)	.602 (.410-.744)	.683 (.524-.802)
	Simplified	.694 (.340-.828)	.753 (.428-.876)	.519 (.315-.672)	.594 (.400-.755)
	Expanded furrows	.739 (.532-.848)	.828 (.658-.913)	.601 (.401-.745)	.692 (.525-.814)
	Fully expanded	.763 (.598-.853)	.824 (.680-.900)	.601 (.394-.747)	.705 (.548-.813)
Proximal	Original	.666 (.452-.801)	.801 (.659-.882)	.472 (.259-.626)	.603 (.417-.734)
	Simplified	.547 (.306-.734)	.700 (.495-.821)	.446 (.242-.603)	.568 (.371-.721)
	Expanded furrows	.694 (.505-.814)	.807 (.678-.884)	.471 (.266-.618)	.585 (.398-.721)
	Fully expanded	.729 (.568-.835)	.791 (.637-.884)	.469 (.255-.620)	.577 (.403-.706)
Distal	Original	.487 (.072-.785)	.730 (.528-.870)	.623 (.428-.754)	.666 (.506-.776)
	Simplified	.560 (.198-.800)	.721 (.497-.863)	.587 (.378-.735)	.665 (.513-.775)
	Expanded furrows	.524 (.154-.774)	.731 (.541-.86)	.623 (.416-.759)	.685 (.507-.797)
	Fully expanded	.503 (.139-.750)	.725 (.553-.844)	.610 (.399-.756)	.702 (.526-.814)
Gastroesophageal junction	Original	.569 (.380-.716)	.580 (.399-.778)	.574 (.361-.722)	.720 (.571-.827)
	Simplified	.516 (.330-.669)	.601 (.417-.779)	.584 (.402-.723)	.728 (.594-.829)
	Expanded furrows	.581 (.394-.725)	.581 (.403-.768)	.572 (.348-.722)	.717 (.560-.83)
	Fully expanded	.598 (.407-.740)	.608 (.425-.784)	.550 (.334-.702)	.719 (.554-.838)
Proximal + distal	Original	.705 (.465-.824)	.787 (.628-.874)	.618 (.413-.759)	.704 (.557-.805)
	Simplified	.68 (.411-.808)	.747 (.528-.852)	.588 (.381-.734)	.663 (.498-.783)
	Expanded furrows	.719 (.494-.826)	.794 (.638-.875)	.625 (.426-.766)	.710 (.559-.815)
	Fully expanded	.716 (.513-.824)	.781 (.627-.861)	.621 (.399-.767)	.722 (.579-.818)

EREFS, Eosinophilic Esophagitis Endoscopic Reference Score.

TABLE 4. Inter- and intrarater reliability of individual endoscopic items

Endoscopic Item		Reliability intraclass correlation coefficient (95% confidence interval)			
		Baseline		Post-treatment	
Feature	Grading definition	Inter-rater	Intrarater	Inter-rater	Intrarater
Exudates	Original (0-2)	.510 (.302-.703)	.775 (.643-.869)	.320 (.062-.561)	.811 (.665-.904)
	Simplified (0-1)	.573 (.372-.735)	.780 (.650-.869)	.344 (.046-.598)	.819 (.686-.891)
	Expanded (0-3)	.408 (.154-.627)	.595 (.332-.792)	.212 (.063-.388)	.714 (.543-.857)
Rings	Original (0-3)	.688 (.498-.811)	.743 (.593-.850)	.633 (.49-.749)	.706 (.563-.811)
	Simplified (0-1)	.611 (.312-.812)	.611 (.316-.812)	.442 (.162-.672)	.589 (.344-.817)
Edema	Original (0-1)	.498 (.179-.662)	.642 (.487-.864)	.339 (.123-.523)	.481 (.295-.687)
	Expanded (0-2)	.419 (.221-.564)	.672 (.535-.782)	.312 (.156-.44)	.564 (.405-.698)
Furrows	Original (0-1)	.536 (.156-.786)	.582 (.255-.874)	.391 (.173-.573)	.536 (.348-.709)
	Expanded (0-2)	.535 (.326-.688)	.686 (.497-.826)	.465 (.225-.614)	.590 (.381-.735)
Stricture	Original (0-1)	.385 (.090-.628)	.674 (.430-.843)	.072 (.000-.244)	.65 (.338-.867)
Inflammatory subscore		.643 (.426-.791)	.814 (.67-.892)	.481 (.22-.664)	.696 (.515-.812)
Fibrostenotic subscore		.748 (.582-.838)	.791 (.642-.885)	.616 (.502-.705)	.715 (.583-.812)

feasibility and operating properties of the EREFS in a multiple component study. We first convened an international expert RAM panel to identify endoscopic items and definitions with face validity and feasibility. These items were used to create modifications of the EREFS, which were then used by blinded central readers to assess pre- and

TABLE 5. Responsiveness and longitudinal validity of the EREFS and its modifications

Item				Weighted correlation, r with	
Segment	EREFS version	Standardized effect size	Area under the receiver-operating characteristic curve*	Change in visual analog scale	Change in peak eosinophil count
Worst disease (global)	Original	1.126 (.757-1.534)	.802 (.717-.867)	.880 (.785-.937)	.194 (–.046-.432)
	Simplified	.974 (.607-1.373)	.765 (.682-.833)	.796 (.631-.896)	.19 (–.052-.436)
	Expanded furrows	1.229 (.858-1.643)	.816 (.726-.882)	.880 (.790-.936)	.177 (–.066-.413)
	Fully expanded	1.252 (.880-1.667)	.806 (.712-.875)	.911 (.849-.949)	.182 (–.066-.423)
Proximal	Original	.654 (.292-1.037)	.680 (.566-.777)	.867 (.754-.93)	.304 (.005-.571)
	Simplified	.695 (.332-1.080)	.691 (.585-.780)	.793 (.628-.89)	.310 (.020-.583)
	Expanded furrows	.683 (.320-1.068)	.687 (.57-.784)	.849 (.735-.917)	.256 (–.049-.532)
	Fully expanded	.697 (.334-1.082)	.689 (.570-.788)	.862 (.751-.929)	.225 (–.097-.511)
Distal	Original	.954 (.587-1.352)	.750 (.628-.842)	.697 (.379-.904)	.044 (–.207-.292)
	Simplified	.852 (.486-1.245)	.734 (.625-.820)	.607 (.258-.874)	.038 (–.221-.294)
	Expanded furrows	1.060 (.692-1.465)	.76 (.631-.855)	0.699 (.375-0.905)	0.063 (–.189-0.317)
	Fully expanded	1.058 (.689-1.462)	.749 (.622-.844)	0.721 (.412-0.917)	0.062 (–.211-0.338)
Gastroesophageal junction	Original	.906 (.540-1.301)	.722 (.608-.812)	.640 (.366-.827)	.245 (–.044-.487)
	Simplified	.581 (.219-0.962)	.659 (.537-.763)	.434 (.146-.678)	.253 (–.048-.52)
	Expanded furrows	.954 (.587-1.352)	.726 (.609-.818)	.635 (.359-.826)	.236 (–.058-.48)
	Fully expanded	.979 (.612-1.378)	.723 (.605-.816)	.655 (.366-.843)	.231 (–.096-.493)
Proximal + distal	Original	.933 (.566-1.329)	.750 (.643-.833)	.886 (.760-.954)	.188 (–.075-.435)
	Simplified	.893 (.527-1.288)	.742 (.646-.819)	.783 (.550-.920)	.188 (–.082-.448)
	Expanded furrows	1.015 (.647-1.416)	.761 (.648-.846)	.876 (.738-.952)	.174 (–.09-.420)
	Fully expanded	1.018 (.651-1.419)	.757 (.642-.844)	.895 (.778-.959)	.158 (–.139-.428)

Values in parentheses are 95% confidence intervals.
EREFS, Eosinophilic Esophagitis Endoscopic Reference Score.
*Area under the receiver-operating characteristic curve represents the nonparametric probability that patients with endoscopic improvement, compared with those without endoscopic improvement, have a lower endoscopic score.

post-treatment disease activity in prospectively collected protocolized endoscopy videos. The EREFS and its modifications, scored in the worst affected area, were associated with moderate-to-substantial inter-rater and moderate-to-almost perfect intrarater reliability. However, EoE-related strictures and stricture diameter were challenging to assess. Although the EREFS and its modifications were highly responsive to treatment, the inflammatory subscore was the most sensitive to change, and there was no benefit to scoring the EREFS by disease location or summing distal and proximal scores. Taken together, our findings suggest that the original major features of the EREFS (exudates, rings, edema, furrows, and stricture) or expanded grading should be used in RCTs as a sensitive measure for detecting treatment effects, with scoring based on the most severe features on global assessment in the esophagus.

Interobserver agreement for assessment of the EREFS components was initially described by Hirano et al[5] using the multirater κ. The authors reported good rater agreement for fixed rings (κ = .50), exudates (κ = .51), furrows (κ = .54), edema (κ = .43), and strictures (κ = .52) when assessed by either experts or nonexperts. In a subsequent study of 30 patients with EoE, van Rhijn et al[25] showed substantial inter-rater agreement for rings (κ = .70) and exudates (κ = .63), moderate agreement for furrows (κ = .49) and strictures (κ = .54), but only

TABLE 6. Responsiveness and longitudinal validity of individual endoscopic items

Item		Standardized effect size (95% CI)	Area under the receiver-operating characteristic curve (95% CI)	Weighted correlation with change in visual analog scale (r [95% CI])
Feature	Grading definition			
Exudates	Original (0-2)	.968 (.600-1.367)	.722 (.603-.816)	.755 (.563-.869)
	Simplified (0-1)	.962 (.595-1.361)	.723 (.604-.817)	.762 (.582-.877)
	Expanded (0-3)	.874 (.508-1.269)	.702 (.600-.787)	.658 (.407-.827)
Rings	Original (0-3)	.504 (.143-.881)	.636 (.477-.770)	.454 (.220-.641)
	Simplified (0-1)	.198 (.164-.566)	.545 (.437-.648)	.210 (–.103-.454)
Edema	Original (0-1)	.524 (.161-.903)	.614 (.542-.682)	.545 (.235-.745)
	Expanded (0-2)	.645 (.282-1.029)	.655 (.556-.742)	.815 (.644-.905)
Furrows	Original (0-1)	.789 (.424-1.179)	.686 (.577-.778)	.693 (.456-.837)
	Expanded (0-2)	1.036 (.667-1.439)	.723 (.592-.825)	.701 (.470-.850)
Stricture	Original (0-1)	.398 (.036-.772)	.562 (.483-.639)	.379 (.126-.590)
Inflammatory subscore		1.217 (.845-1.631)	.787 (.661-.875)	.782 (.606-.885)
Fibrostenotic subscore		.558 (.196-.938)	.644 (.48-.779)	.461 (.245-.642)

CI, Confidence interval.

slight agreement for edema ($\kappa = .12$) when assessed from still images. Still images may improve the reliability of stricture assessment based on static evaluation of absolute luminal diameter, whereas assessment of vascularity may be better appreciated on a dynamic video with esophageal insufflation. In the current study, expert central readers reliably assessed the EREFS and its component items, except for strictures in upper endoscopy video recordings. An analogous situation has been observed in patients with Crohn's disease, when similar methodology demonstrated poor reliability for the evaluation of ileal or colonic strictures.[26] There are several possible explanations for this observation. First, dynamic stricture detection depends on the degree of insufflation and force attempted for endoscope passage, yet central readers are unable to gauge procedural tactile feedback. Second, endoscopy is not the ideal modality to measure esophageal diameter compared with radiographic assessment[27] or endoluminal functional impedance planimetry.[28] Third, uniform criteria do not exist for the definition of an esophageal stricture greater than the caliber of the endoscope but less than normal esophageal diameter. For instance, a 15-mm luminal diameter is abnormal but not considered a clinically relevant stricture by most endoscopists. Given that stricture development is an important sequela of untreated EoE with substantial contribution to patient symptoms,[29,30] further evaluation of the reliability of this item is required before it can be confidently excluded from endoscopic scoring systems.

We evaluated several modifications of the EREFS based on RAM outcomes. We hypothesized that expanded grading definitions and evaluation of multiple esophageal segments would better capture disease extent and improve responsiveness. The expanded versions of the EREFS were numerically most responsive, although all versions showed large effect sizes and adequate longitudinal validity with a change in VAS. Responsiveness was not improved with separate scoring or summation of the proximal and distal esophagus. Similar effect sizes for the EREFS have been observed in cohort studies and RCTs.[7,31-35] For example, the mean difference in the EREFS score after 12 weeks was –3.8 (SD, 3.9) in the treatment arm compared with +.4 (SD, 6.7) in the placebo arm ($P < .0001$) in a phase II trial of budesonide oral suspension compared with placebo.[32] In the EOS-1 phase III induction trial, patients randomized to a budesonide orodispersible tablet had a mean change in the EREFS score of –2.6 (95% CI, –3.1 to –2.1) compared with –.1 (95% CI, –.8 to +.5) for patients randomized to placebo ($P < .0001$).[35] These large effect sizes likely reflect the significant efficacy of topical corticosteroids in EoE, although similar changes have also been demonstrated for therapies with other mechanisms of action, including interleukin-4 receptor and interleukin-13 blockade.[33,34]

The inflammatory subscore was the most responsive component of the EREFS, again likely because of the treatment modalities evaluated in this study, which all aim to reduce the burden of eosinophilic inflammation. The short treatment period typical of induction trials may also explain this observation (when compared with the fibrostenotic subscore), which may be less pronounced in a maintenance trial. Our findings are similar to those observed in the EOS-1 trial, where most of the change in the EREFS in the active treatment group was driven by reductions in the inflammatory subscore, because there was no significant difference in the fibrotic subscore compared with

placebo (–.4 in the treatment group vs –.1 in the placebo group, $P = .22$).[35] Antifibrotic therapies are nevertheless needed, and development of therapies that target esophageal remodeling may be effective at reducing endoscopic strictures or fixed rings.[36] Furthermore, fibrostenotic consequences of EoE are associated with clinically relevant outcomes and the need for esophageal dilation and thus should be assessed.

Other modifications to the EREFS were considered by the RAM panel. Dellon et al[7] suggested increasing the weighting of the inflammatory components of the EREFS subscore to improve responsiveness. In a prospective cohort of 67 incident cases of EoE treated with topical swallowed corticosteroids or dietary elimination, doubling the weighting for exudates, rings, and edema maximized the responsiveness of the total EREFS score (3.19 change in weighted vs 1.87 unweighted EREFS). In contrast, Schoepfer et al[8] evaluated multiple methods of scoring the EREFS that were regressed against a global assessment of endoscopic disease activity and tested for responsiveness in a phase Ib/IIa RCT including patients treated with fluticasone (n = 16) or placebo (n = 8). No significant benefit to weighting individual items such as exudates was observed, particularly because scoring of exudates contributed most to interendoscopist variation. Similarly, there was no significant benefit to scoring proximal or distal findings separately, and the authors concluded that the variations were not superior to the original EREFS. Although numerical estimates of reliability and responsiveness were greatest when the worst affected area was scored, this method may not be appropriate for locally acting therapies as compared with systemic agents. Although a simplified, dichotomized method for scoring of the EREFS may be more feasible in RCTs, particularly when endoscopic outcomes are evaluated locally by multiple investigators across multiple sites, as is the current practice, this method was not more reliable, and estimates of responsiveness were numerically lower compared with the original and expanded EREFS versions explored in this study.

Our study has several strengths. We used a rigorous, multiple-step method to assess the performance metrics of endoscopic evaluation in EoE. High-quality videos before and after treatment were evaluated in duplicate by expert blinded central readers, resulting in over 320 observations for analysis. Estimation of reliability and responsiveness by time point, disease location, and individual endoscopic feature provided a comprehensive assessment of reader performance. We also acknowledge some important limitations. First, responsiveness is ideally evaluated using a dataset from an RCT of a therapy of known efficacy using treatment assignment as the criterion for change. Although these data were not readily available, we obtained high-quality endoscopy recordings from patients managed in a protocolized fashion at 2 centers, and the criterion for change used in this study (the SD of the VAS) is a well-accepted benchmark. Second, the study was powered to evaluate the minimum acceptable reliability between readers. However, a larger sample size would be needed to evaluate differences in reliability and responsiveness by treatment assignment (proton pump inhibitor vs topical corticosteroids vs dietary exclusion). Third, we used 2 central readers in this study, recognizing that the precision of the ICC estimates would have been improved with more readers. However, we designed this study to mimic an RCT setting, using protocolized, prospective trial data with 2 blinded readers, evaluating videos on a central image management solution currently used for RCTs. In an RCT setting, it is unlikely that more than 2 central readers would be used because of considerations of cost and feasibility. The central readers in this study are highly experienced in the use of the EREFS, which may limit the generalizability of these results to other readers or local site endoscopists in a clinical trial. However, there are important advantages to centralized reading for minimizing observation bias, and this has become the standard for assessment of endoscopic endpoints in other areas. Fourth, the results of this study are not generalizable to pediatric patients, a population with substantial unmet medical need.[31] Finally, we did not formally test comparisons between segments or calculation methods because this is statistically infeasible with >2000 different possible permutations that would be subject to a very high false-positive detection rate and includes comparisons between segments and calculation methods that have limited clinical applicability.

In conclusion, our findings improve the current use of the EREFS. The EREFS and its modifications are reliably assessed by expert central readers and are highly responsive to anti-inflammatory therapy. Our study highlights that centralized reading of videos for endoscopic endpoints can be considered in EoE RCTs to minimize the risk of bias and that either the original or expanded versions of the EREFS should be scored based on the worst affected area to maximize the likelihood of detecting treatment effects. Future research should focus on improving the endoscopic assessment of strictures in EoE RCTs and determining whether fibrostenotic features are responsive to other pharmacologic mechanisms of action.

ACKNOWLEDGMENTS

We acknowledge Ms Lee Anne Williamson for her role as the project manager for this work and Dr. Willemijn E. de Rooij for providing clinical data. All deidentified data from this analysis is available on request.

REFERENCES

1. de Rooij WE, Dellon ES, Parker CE, et al. Pharmacotherapies for the treatment of eosinophilic esophagitis: state of the art review. Drugs 2019;79:1419-34.

2. Lyons E, Donohue K, Lee JJ. Developing pharmacologic treatments for eosinophilic esophagitis: draft guidance from the United States Food and Drug Administration. Gastroenterology 2019;157:275-7.
3. Ma C, Schoepfer AM, Dellon ES, et al. Development of a core outcome set for therapeutic studies in eosinophilic esophagitis (COREOS). J Allergy Clin Immunol 2022;149:659-70.
4. Kim HP, Vance RB, Shaheen NJ, et al. The prevalence and diagnostic utility of endoscopic features of eosinophilic esophagitis: a meta-analysis. Clin Gastroenterol Hepatol 2012;10:988-96.
5. Hirano I, Moy N, Heckman MG, et al. Endoscopic assessment of the oesophageal features of eosinophilic oesophagitis: validation of a novel classification and grading system. Gut 2013;62:489-95.
6. Ma C, van Rhijn BD, Jairath V, et al. Heterogeneity in clinical, endoscopic, and histologic outcome measures and placebo response rates in clinical trials of eosinophilic esophagitis: a systematic review. Clin Gastroenterol Hepatol 2018;16:1714-29.
7. Dellon ES, Cotton CC, Gebhart JH, et al. Accuracy of the eosinophilic esophagitis endoscopic reference score in diagnosis and determining response to treatment. Clin Gastroenterol Hepatol 2016;14:31-9.
8. Schoepfer AM, Hirano I, Coslovsky M, et al. Variation in endoscopic activity assessment and endoscopy score validation in adults with eosinophilic esophagitis. Clin Gastroenterol Hepatol 2019;17:1477-88.
9. Fitch K, Bernstein S, Aguilar M, et al. The RAND/UCLA appropriateness method user's manual. Santa Monica, CA: RAND Corporation; 2001.
10. Warners MJ, Hindryckx P, Levesque BG, et al. Systematic review: disease activity indices in eosinophilic esophagitis. Am J Gastroenterol 2017;112:1658-69.
11. U.S. Food and Drug Administration. Center for Drug Evaluation and Research: clinical outcome assessment (COA) compendium. 2021. Available at: https://www.fda.gov/drugs/development-resources/clinical-outcome-assessment-compendium. Accessed February 27, 2021.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
13. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measure 1973;33:613-9.
14. Berk RA. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. Am J Ment Defic 1979;83:460-72.
15. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582-92.
16. Marchal-Bressenot A, Salleron J, Boulagnon-Rombi C, et al. Development and validation of the Nancy histological index for UC. Gut 2017;66:43-9.
17. Mosli MH, Feagan BG, Zou G, et al. Development and validation of a histological index for UC. Gut 2017;66:50-8.
18. Hindryckx P, Jairath V, Zou G, et al. Development and validation of a magnetic resonance index for assessing fistulas in patients with Crohn's disease. Gastroenterology 2019;157:1233-44.
19. Jairath V, Ordas I, Zou G, et al. Reliability of measuring ileo-colonic disease activity in Crohn's disease by magnetic resonance enterography. Inflamm Bowel Dis 2018;24:440-9.
20. Cohen J. A power primer. Psychol Bull 1992;112:155-9.
21. Safroneeva E, Straumann A, Coslovsky M, et al. Symptoms have modest accuracy in detecting endoscopic and histologic remission in adults with eosinophilic esophagitis. Gastroenterology 2016;150:581-90.
22. Ma C, Schoepfer AM, Safroneeva E, et al. Development of a Core Outcome Set for Therapeutic Studies in Eosinophilic Esophagitis (CORE-EOS): an international multidisciplinary consensus. Gastroenterology 2021;161:748-55.
23. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012;31:3972-81.
24. Ishimura N, Sumi S, Okada M, et al. Ankylosaurus back sign: novel endoscopic finding in esophageal eosinophilia patients indicating proton pump inhibitor response. Endosc Int Open 2018;6:E165-72.
25. van Rhijn BD, Warners MJ, Curvers WL, et al. Evaluating the endoscopic reference score for eosinophilic esophagitis: moderate to substantial intra- and interobserver reliability. Endoscopy 2014;46:1049-55.
26. Khanna R, Zou G, D'Haens G, et al. Reliability among central readers in the evaluation of endoscopic findings from patients with Crohn's disease. Gut 2016;65:1119-25.
27. Lee J, Huprich J, Kujath C, et al. Esophageal diameter is decreased in some patients with eosinophilic esophagitis and might increase with topical corticosteroid therapy. Clin Gastroenterol Hepatol 2012;10:481-6.
28. Read AJ, Pandolfino JE. Biomechanics of esophageal function in eosinophilic esophagitis. J Neurogastroenterol Motil 2012;18:357-64.
29. Schoepfer AM, Straumann A, Panczak R, et al. Development and validation of a symptom-based activity index for adults with eosinophilic esophagitis. Gastroenterology 2014;147:1255-66.
30. Schoepfer AM, Safroneeva E, Bussmann C, et al. Delay in diagnosis of eosinophilic esophagitis increases risk for stricture formation in a time-dependent manner. Gastroenterology 2013;145:1230-6.
31. Wechsler JB, Bolton SM, Amsden K, et al. Eosinophilic esophagitis reference score accurately identifies disease activity and treatment effects in children. Clin Gastroenterol Hepatol 2018;16:1056-63.
32. Dellon ES, Katzka DA, Collins MH, et al. Budesonide oral suspension improves symptomatic, endoscopic, and histologic parameters compared with placebo in patients with eosinophilic esophagitis. Gastroenterology 2017;152:776-86.
33. Hirano I, Collins MH, Assouline-Dayan Y, et al. RPC4046, a monoclonal antibody against IL13, reduces histologic and endoscopic activity in patients with eosinophilic esophagitis. Gastroenterology 2019;156:592-603.
34. Hirano I, Dellon ES, Hamilton JD, et al. Efficacy of dupilumab in a phase 2 randomized trial of adults with active eosinophilic esophagitis. Gastroenterology 2020;158:111-22.
35. Lucendo AJ, Miehlke S, Schlag C, et al. Efficacy of budesonide orodispersible tablets as induction therapy for eosinophilic esophagitis in a randomized placebo-controlled trial. Gastroenterology 2019;157:74-86.
36. Nhu QM, Hsieh L, Dohil L, et al. Antifibrotic effects of the thiazolidinediones in eosinophilic esophagitis pathologic remodeling: a preclinical evaluation. Clin Transl Gastroenterol 2020;11:e00164.

Abbreviations: AUC, area under the receiver-operating characteristic curve; EoE, eosinophilic esophagitis; EREFS, Eosinophilic Esophagitis Endoscopic Reference Score; ICC, intraclass correlation coefficient; RAM, Research and Development University of California Los Angeles appropriateness methodology; RCT, randomized controlled trial; SD, standard deviation; SES, standardized effect size; VAS, visual analog scale.

DISCLOSURE: *The following authors disclosed financial relationships: C. Ma: Consultant for AVIR Pharma Inc and Alimentiv Inc; speaker for AVIR Pharma Inc. A. J. Bredenoord: Research support from Norgine, Nutricia, Thelial, SST, and Bayer; speaker and consultant for AstraZeneca, Medtronic, Laborie, Alimentiv, Celgene, DrFalkPharma, Arena, Esocap, Calypso, and Regeneron. E. S. Dellen: Research support from Adare/Ellodi, Allakos, AstraZeneca, Banner, Holoclara. GlaxoSmithKline, Meritage, Miraca, Nutricia, Celgene/Receptos/BMS, Regeneron, and Shire/Takeda; consultant for Abbott, Abbvie, Adare/Ellodi, Aimmune, Allakos, Amgen, Arena, AstraZeneca, Avir, Biorasi, Calypso, Celgene/Receptos/BMS, Celldex, Eli Lilly, EsoCap, GlaxoSmithKline, Gossamer Bio, Landos, Morphic, Parexel/Calyx, Regeneron, Robarts/Alimentiv Inc, Salix, Sanofi, and Shire/Takeda. J. A. Alexander: Received research support from Regeneron, Adare, Arena, Celegene, and Ellode; consultant for Alimentiv; financial interest in Meritage Pharmacia. L. B Biedermann: Consultant and/or Falk Foundation, Vifor AG, Esocap AG, Sanofi-Aventis AG, and Calypso Biotech SA. M. Hogan, L. Guizzetti, J. Rémillard: Employee of Alimentiv Inc. G. Zou, L. M. Shackelton: Consultant for Alimentiv Inc. D. A. Katzka: Consultant for Takeda and Celgene. M. Chehade: Research*

support from Regeneron, Allakos, Shire, AstraZeneca, and Danone; consultant for Regeneron, Allakos, Adare, Shire/Takeda, AstraZeneca, Sanofi, Ellodi, Phathom, and Bristol Myers Squibb. G. W. Falk: Research support from Adare/Ellodi, Arena, Allakos, Celgene/Receptos/BMS, Lucid, Regeneron, and Shire/Takeda; consultant for Adare/Ellodi, Allakos, Arena, Celgene/Receptos/BMS, Lucid, Regeneron, Sanofi, and Shire/Takeda. G. T. Furuta: Research support from Holoclara and Arena; consultant for Takeda; cofounder of EnteroTrack. S. K. Gupta: Consultant for Abbott, Adare, Allakos, Celgene, Gossamer Bio, QOL, UpToDate, Medscape, and Viaskin; research support from Shire, Allakos, Adare, and National Institutes of Health U54 grant to the Consortium of Eosinophilic Gastrointestinal Disease Researchers. A. M. Schoepfer: Research support from Adare/Ellodi, AstraZeneca, Receptos/BMS, Dr Falk Pharma, Regeneron, and Vifor Pharma; consultant for Adare/Ellodi, Amgen, AstraZeneca, Celgene/Receptos/BMS, Dr Falk Pharma, GlaxoSmithKline, Gossamer Bio, Regeneron, Sanofi-Genzyme, and Vifor Pharma. S. Miehlke: Consultant for Dr Falk Pharma, EsoCap, Sanofi-Regeneron, and Celgene; Falk Foundation/Falk Foundation. F. J. Moawad: Consultant for Medtronic, Shire/Takeda, and Sanofi. K. Peterson: Research support from Adare/Ellodi, Allakos, AstraZeneca, Receptos/BMS, Regeneron, Shire/Takeda, and Chobani; consultant for Adare/Ellodi, Allakos, Alladapt, AstraZeneca, Celgene/Receptos/BMS, Regeneron, Sanofi, and Shire/Takeda. N. P. Gonsalves: Consultant for Allakos, Sanofi-Regeneron, Abbvie, Astra-Zeneca, and Nutriticia; royalties from Up-to-Date. A. Straumann: Consultant for Allakos, Astra-Zeneca, Calypso, EsoCap, Falk Pharma, Gossamer, Nutricia, Pfizer, Receptos-Celgene, Regeneron-Sanofi, Roche-Genentec, Shire, and Tillotts. J. B. Wechsler: Consultant for Allakos, Regeneron, and Sanofi/Genzyme. B. G. Feagan: Consultant for AbbVie, AdMIRx, AgomAB Therapeutics, Akebia, Alivio Therapeutics, Allakos, Amgen, Applied Molecular Transport Inc, Arena Pharma, Avir, Azora Therapeutics, Boehringer-Ingelheim, Boston Pharma, Celgene/BMS, Connect BioPharma, Cytoki, Disc Medicine, Everest Clinical Research Corp, Eli Lilly, Equillium, Ferring, Galapagos, Galen Atlantica, Genentech/Roche, Gilead, Glenmark, Gossamer Pharma, GSK, Hoffmann-LaRoche, Hot Spot Therapeutics, Index Pharma, ImmunExt, Imbotex, Intact Therapeutics, Janssen, Japan Tobacco Inc, Kaleido Biosciences, Leadiant, Millennium, MiroBio, Morphic Therapeutics, Mylan, OM Pharma, Origo BioPharma, Otsuka, Pandion Therapeutics, Pfizer, Prometheus Therapeutics and Diagnostics, Progenity, PTM Therapeutics, Q32 Bio, Rebiotix, RedHill, Biopharma, REDX, Sandoz, Sanofi, Seres Therapeutics, Surrozen Inc, Takeda, Teva, Thelium, Theravance, Tigenix, Tillotts, UCB Pharma, VHsquared Ltd, Viatris, Ysios, and Zealand Pharma; Speaker's Bureau for AbbVie, Janssen, and Takeda; Member, Scientific Advisory Board for AbbVie, Amgen, Boehringer-Ingelheim, Celgene/BMS Genentech/Roche, Janssen, Novartis, Origo BioPharma, Pfizer, Prometheus, Takeda, Tillotts Pharma, Teva, Progenity, Index, Ecor1Capital, Morphic, and GSK; Stock shareholder in Gossamer Pharma; Employee, Western University and Alimentiv Inc. V. Jairath: Consultant for AbbVie, Alimentiv Inc (formerly Robarts Clinical Trials), Arena Pharmaceuticals, Bristol Myers Squibb, Celltrion, Eli Lilly, Ferring, Fresenius Kabi, GlaxoSmithKline, Genetech, Gilead, Janssen, Merck, Mylan, Pendopharm, Pfizer, Roche, Sandoz, Takeda, and Topivert; speaker for Abbvie, Ferring, Janssen Pfizer Shire, and Takeda. I. Hurano: Research support from Adare/Ellodi, Allakos, AstraZeneca, Meritage, Receptos/BMS, Regeneron, and Shire/Takeda; consultant for Adare/Ellodi, Allakos, Amgen, Arena, AstraZeneca, Calypso, Celgene/Receptos/BMS, Eli Lilly, EsoCap, GlaxoSmithKline, Gossamer Bio, Parexel, Regeneron, Sanofi, and Shire/Takeda. All other authors disclosed no financial relationships.

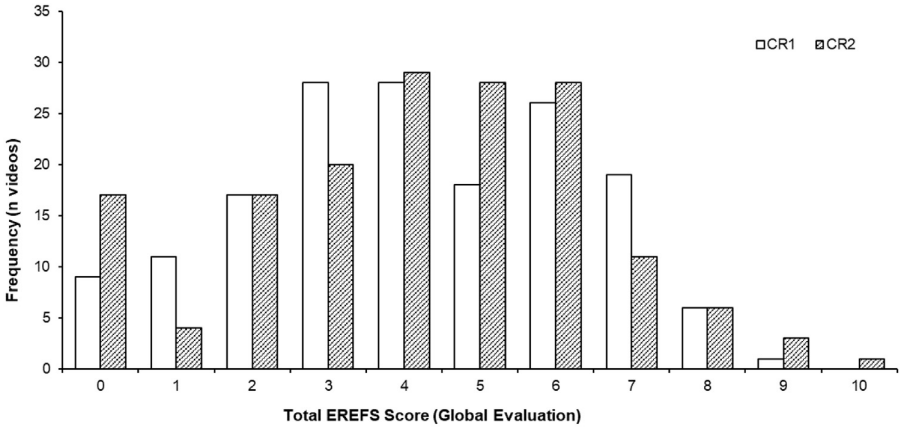
0016-5107/$36.00
https://doi.org/10.1016/j.gie.2022.01.014

Received August 30, 2021. Accepted January 21, 2022.

Current affiliations: Division of Gastroenterology & Hepatology, Departments of Medicine and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada (1), Alimentiv Inc, London, Ontario, Canada (2), Department of Gastroenterology & Hepatology, Academic Medical Center, Amsterdam, The Netherlands (3), Division of Gastroenterology and Hepatology, Department of Medicine, Center for Esophageal Diseases and Swallowing, Center for Gastrointestinal Biology and Disease, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA (4), Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, USA (5), Department of Gastroenterology and Hepatology (6), Swiss EoE Clinics and Research Network, Department of Gastroenterology and Hepatology (19), University Hospital Zurich, Zurich, Switzerland; Department of Epidemiology and Biostatistics (7), Division of Gastroenterology and Hepatology (20), Western University, London, Ontario, Canada; Mount Sinai Center for Eosinophilic Disorders, Icahn School of Medicine at Mount Sinai, New York, New York, USA (8), Division of Gastroenterology and Hepatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA (9), Digestive Health Institute, Children's Hospital Colorado, Gastrointestinal Eosinophilic Diseases Program, University of Colorado School of Medicine, Aurora, Colorado, USA (10), Division of Pediatric Gastroenterology, Hepatology and Nutrition, Riley Children's Hospital, Indiana University School of Medicine, Community Health Network, Indianapolis, Indiana, USA (11), Division of Gastroenterology, Hepatology & Nutrition, Department of Pediatrics, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, USA (12), Division of Pediatric Gastroenterology, Department of Pediatrics, John H. Stroger Hospital of Cook County, Chicago, Illinois, USA (13), Division of Gastroenterology and Hepatology, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland (14), Center for Esophageal Diseases, University Hospital Hamburg-Eppendorf, Hamburg, Germany (15), Division of Gastroenterology, Scripps Clinic, La Jolla, California, USA (16), Division of Gastroenterology and Hepatology, University of Utah, Salt Lake City, Utah, USA (17), Division of Gastroenterology and Hepatology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA (18).

Reprint requests: Christopher Ma, MD MPH, Division of Gastroenterology and Hepatology, University of Calgary, 6D61, Teaching Research Wellness Bldg, 3280 Hospital Dr NW, Calgary, Alberta, Canada, T2N 3V9.

APPENDIX



Supplementary Figure 1. Distribution of scoring for the Eosinophilic Esophagitis Endoscopic Reference Score (EREFS) by central reader (CR).

SUPPLEMENTARY TABLE 1. Original Eosinophilic Esophagitis Endoscopic Reference Score (excluding minor feature of crepe paper esophagus)

Major features	Grading
Fixed rings (also referred to as concentric rings, corrugated esophagus, corrugated rings, ringed esophagus, trachealization)	Grade 0: none Grade 1: mild (subtle circumferential ridges) Grade 2: moderate (distinct rings that do not impair passage of a standard diagnostic adult endoscope [outer diameter, 8-9.5 mm]) Grade 3: severe (distinct rings that do not permit passage of a diagnostic endoscope)
Exudates (also referred to as white spots, plaques)	Grade 0: none Grade 1: mild (lesions involving <10% of the esophageal surface area) Grade 2: severe (lesions involving >10% of the esophageal surface area)
Furrows (also referred to as vertical lines, longitudinal furrows)	Grade 0: absent Grade 1: present
Edema (also referred to as decreased vascular markings, mucosal pallor)	Grade 0: absent (distinct vascularity present) Grade 1: loss of clarity or absence of vascular markings
Stricture	Grade 0: absent Grade 1: present

SUPPLEMENTARY TABLE 2. Baseline patient demographic characteristics

Characteristic	Value
Age, y	
Mean (SD)	41.3 (12.7)
Median (IQR)	39.5 (31-47.5)
Gender	
Female	15 (37.5)
Male	25 (62.5)
Peak eosinophil count at baseline	
Mean (SD)	61.8 (23.3)
Median (IQR)	50 (50-80)
Peak eosinophil count at follow-up	
Mean (SD)	12.6 (20.4)
Median (IQR)	2 (0-20)
Race	
White	39 (97.5)
Asian	1 (2.5)
Concurrent atopic disease	
Yes	34 (85)
No	6 (15)

Values are n (%) unless otherwise defined.
SD, Standard deviation; *IQR*, interquartile range.