RESEARCH CORRESPONDENCE

Reliability of Endoscopic Evaluation of Postoperative Recurrent Crohn's Disease

Q9 Christopher Ma,^{*,‡,a} Krisztina B. Gecse,^{§,a} Marjolijn Duijvestein,^{§,a}
 William J. Sandborn,^{‡,||} Guangyong Zou,^{‡,||} Lisa M. Shackleton,[‡] Larry W. Stitt,[‡]
 Claire E. Parker,[‡] Peter Bossuyt,^{#,**} Mark Löwenberg,[§] Reena Khanna,^{‡‡}
 Severine Vermeire,^{**} Paul Rutgeerts,^{**} Brian G. Feagan,^{‡,||,**} Vipul Jairath,^{‡,||,**} and
 Q4 Geert R. D'Haens^{‡,§}

^{*}Division of Gastroenterology and Hepatology, University of Calgary, Calgary, Canada; [‡]Robarts Clinical Trials, London, Canada; [§]Department of Gastroenterology and Hepatology, Amsterdam UMC, University Medical Center, Amsterdam, the Netherlands; ^{||}Division of Gastroenterology, University of California, San Diego, La Jolla, California; ^{||}Department of Epidemiology and Biostatistics, Western University, London, Canada; [#]Imelda Gastrointestinal Clinical Research Center, Imelda General Hospital, Bonheiden, Belgium; ^{**}Division of Gastroenterology and Hepatology, University Hospitals Leuven, Leuven, Belgium; and ^{‡‡}Division of Gastroenterology, Western University, London, Canada

Endoscopic evaluation for postoperative recurrence of Crohn's disease (CD) is routinely integrated into clinical practice. The Rutgeerts score (RS) was developed to grade the severity of endoscopic postoperative CD recurrence and has been integrated into clinical practice guidelines and utilized as an endpoint in interventional trials.^{1,2} However, the operating properties of the RS have not been fully assessed. Furthermore, the RS i2 grade groups purely anastomotic ulcerations with those in the neoterminal ileum, whereas the modified Endoscopic Postoperative Recurrence Score (mEPRS) distinguishes lesions limited to the ileocolic anastomosis (i2a) from those in the neoterminal ileum (i2b). Accurate characterization of endoscopic recurrence is an important determinant for initiating postoperative medical therapy. Therefore, variability in endoscopic scoring may result in inappropriate therapeutic decisions.³ We evaluated the reliability of endoscopic assessment of postoperative CD recurrence among independent blinded central readers.

Materials and Methods

Video recordings from 50 consecutive CD patients with a history of ileocolic resection undergoing ileocolonoscopy at 2 European hospitals were independently and randomly assessed in triplicate by 4 experienced gastroenterologists blinded to all clinical information (599 total assessments; 50 videos × 4 readers × 3 assessments; 1 missing assessment).

Mucosal appearance was evaluated using the RS, mEPRS, and a global evaluation of lesion severity (GELS) measured on a 10-cm linear analog scale (Supplemental Methods). Inter- and intrarater reliability was quantified by the intraclass correlation coefficient (ICC). The

strength of agreement was interpreted according to the criteria established by Landis and Koch.⁴

We calculated the degree of misclassification arising from interrater disagreement that would result in inappropriate therapeutic initiation assuming treatment is required for patients with RS >i1. Under a binomial distribution, we calculated the probability that a single evaluation would lead to inappropriate treatment based on the number of cases in which disagreement was observed.

Results

The distribution of mEPRS scores was 21% (125 of 599) i0, 8% (49 of 599) i1, 17% (103 of 599) i2a, 23% (136 of 599) i2b, 8% (46 of 599) i3, and 23% (140 of 599) i4.

Interrater ICCs for the RS, mEPRS, and GELS scores were 0.71 (95% confidence interval [CI], 0.59–0.80), 0.74 (95% CI, 0.63–0.82), and 0.73 (95% CI, 0.62–0.81), respectively. Intrarater ICCs for the RS, mEPRS, and GELS scores were 0.85 (95% CI, 0.78–0.89), 0.86 (95% CI, 0.80–0.90), and 0.81 (95% CI, 0.71–0.87), respectively. ICCs for individual component items are summarized in Table 1. The ICC estimates for intrarater reliability were higher than those observed for interrater reliability for all individual component items. The correlation coefficient between the RS and GELS was 0.77 (95% CI, 0.69–0.82) and 0.78 (95% CI, 0.71–0.83) between the mEPRS and GELS.

^aAuthors share co-first authorship.

Table 1. Reliability of Endoscopic Assessment for Postoperative Crohn's Disease Recurrence

Score/Item	Reliability (ICC With 95% CI)	
	Interrater	Intrater
Rutgeerts score (EPRS)	0.71 (0.59–0.80)	0.85 (0.78–0.89)
mEPRS	0.74 (0.63–0.82)	0.86 (0.80–0.90)
GELS	0.73 (0.62–0.81)	0.81 (0.71–0.87)
Anastomotic lesions		
Aphthous ulcer(s) <5 mm	0.12 (0.02–0.21)	0.48 (0.30–0.62)
Large ulcer(s) ≥5 mm	0.37 (0.24–0.49)	0.60 (0.52–0.69)
Any stenosis	0.49 (0.23–0.69)	0.81 (0.65–0.91)
Neoterminal ileum lesions		
Aphthous ulcer(s) <5 mm	0.17 (0.04–0.27)	0.48 (0.25–0.64)
Large ulcer(s) ≥5 mm	0.40 (0.24–0.55)	0.60 (0.48–0.71)
Any stenosis	0.40 (0.06–0.65)	0.80 (0.62–0.91)

CI, confidence interval; EPRS, Endoscopic Postoperative Recurrence Score; GELS, global evaluation of lesion severity; ICC, intraclass correlation coefficient; mEPRS, modified Endoscopic Postoperative Recurrence Score.

Assuming the majority of readers made the correct evaluation for the treatment thresholds of RS ≥i2, mEPRS ≥i2a, and mEPRS ≥i2b, the probability of inappropriate therapeutic initiation from a single read was 9.0% (95% CI, 2.5%–15.5%), 11.5% (95% CI, 4.4%–18.6%), and 11.5% (95% CI, 4.4%–18.6%), respectively.

Discussion

Accurate postoperative risk stratification for patients with CD recurrence is dependent on reliable identification of high-risk lesions at endoscopy. Here, we demonstrate that the RS and mEPRS are reliable instruments with “substantial” to “almost perfect” inter- and intrater reliability when evaluated by expert central readers. However, approximately 10% of therapeutic decisions made based on thresholds of RS ≥i2 or mEPRS ≥i2a or i2b could still be inappropriate.

Kennedy et al⁵ previously demonstrated a similar interrater ICC for evaluation of still images in the TOPPIC trial (0.82; 95% CI, 0.74–0.88). Daperno et al⁶ showed “fair” (Fleiss' $\kappa = 0.57$) to “good” reliability ($\kappa = 0.67$), although only 10 videos were assessed. Marteau et al⁷ concluded there was “moderate” interrater reliability (unweighted $\kappa = 0.43$; 95% CI, 0.33–0.52) in two randomized trials of patients with postoperative CD. Inconsistency in strength of association compared to our study may be due to differences in statistical methodology. Compared to the unweighted kappa, ICCs give more weight to extreme disagreements that are more clinically relevant on a 5-point ordinal scale. Although we found the overall reliability for the RS and mEPRS to be substantial, the interrater reliability for aphthous ulcerations at the anastomosis or neoterminal ileum was poor, potentially reflecting ambiguity in the demarcation between anastomosis and ileum and distinguishing superficial small ulcerations from residual mucous or debris.⁸

This study captured a representative spectrum of postsurgical endoscopic CD severity, included triplicate assessments by multiple blinded central readers, and evaluated the mEPRS that may differentiate treatment response between anastomotic and ileal ulcerations. However, generalizability to nonexpert endoscopists is unclear. Additional research is required to assess the responsiveness of these indices to treatment and to validate their use either alone, or in combination with radiography and biomarkers for prediction of long-term outcomes.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Clinical Gastroenterology and Hepatology* at www.cghjournal.org, and at <https://doi.org/10.1016/j.cgh.2019.08.046>.

References

1. Rutgeerts P, Geboes K, Vantrappen G, et al. Predictability of the postoperative course of Crohn's disease. *Gastroenterology* 1990;99:956–963.
2. Nguyen GC, Loftus EV Jr, Hirano I, et al. American Gastroenterological Association Institute guideline on the management of Crohn's disease after surgical resection. *Gastroenterology* 2017;152:271–275.
3. Gionchetti P, Dignass A, Danese S, et al. 3rd European evidence-based consensus on the diagnosis and management of Crohn's disease 2016: part 2: surgical management and special situations. *J Crohns Colitis* 2017;11:135–149.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
5. Kennedy NA, Ennis H, Gaya DR, et al. Su1186 Interobserver agreement in assessment of Rutgeerts' score of endoscopic recurrence of ileal Crohn's disease. *Gastroenterology* 2015; 148:S-431.
6. Daperno M, Comberlato M, Bossa F, et al. Inter-observer agreement in endoscopic scoring systems: preliminary report of an ongoing study from the Italian Group for Inflammatory Bowel Disease (IG-IBD). *Dig Liver Dis* 2014;46:969–973.
7. Marteau P, Laharie D, Colombel JF, et al. Interobserver variation study of the Rutgeerts Score to assess endoscopic recurrence after surgery for Crohn's disease. *J Crohns Colitis* 2016; 10:1001–1005.
8. 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;65:1119–1125.

Reprint requests

Address requests for reprints to: Geert R. D'Haens, Department of Gastroenterology & Hepatology, Amsterdam UMC, University Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands. e-mail: g.dhaens@amc.uva.nl; fax: ●●●.

Conflicts of interest

These authors disclose the following: Christopher Ma reports consulting fees from Robarts Clinical Trials, Janssen, and AbbVie; and speaker's fees from Janssen and Pfizer. Krisztina B. Gecse reports having served as speaker and/or advisor for Amgen, AbbVie, Boehringer Ingelheim, Ferring, Hospira, MSD, Pfizer, Samsung Bioepis, Sandoz, Takeda, and Tigenix. Marjolijn Duijvestein reports having served as an advisor for Echo pharma and Robarts Clinical Trials, Inc; nonfinancial support from Dr Falk Pharma; and speaker's fees from Janssen, Merck & Co., Inc, Pfizer, Takeda, and Tillotts Pharma. William J.

Sandborn reports research grants from Atlantic Healthcare Limited, Amgen, Genentech, Gilead Sciences, AbbVie, Janssen, Takeda, Lilly, Celgene/Receptos, Pfizer, Prometheus Laboratories (owned by Precision IBD); consulting fees from AbbVie, Allergan, Amgen, Arena Pharmaceuticals, Avexigen Therapeutics, BeiGene, Boehringer Ingelheim, Celgene, Celltrion, Conatus, Cosmo, Escalier Biosciences, Ferring, Forbion, Genentech, Gilead Sciences, Gossamer Bio, Incyte, Janssen, Kyowa Kirin Pharmaceutical Research, Landos Biopharma, Lilly, Oppilan Pharma, Otsuka, Pfizer, Precision IBD (owns Prometheus Laboratories), Progenity, Prometheus Laboratories (owned by Precision IBD), Reistone, Ritter Pharmaceuticals, Roberts Clinical Trials, Inc (owned by Health Academic Research Trust, HART), Series Therapeutics, Shire, Sienna Biopharmaceuticals, Sigmoid Biotechnologies, Sterna Biologicals, Sublimity Therapeutics, Takeda, Theravance Biopharma, Tigenix, Tillotts Pharma, UCB Pharma, Ventyx Biosciences, Vimalan Biosciences, and Vivelix Pharmaceuticals; stock ownership or stock options from BeiGene, Escalier Biosciences, Gossamer Bio, Oppilan Pharma, Precision IBD (owns Prometheus Laboratories), Progenity, Ritter Pharmaceuticals, Ventyx Biosciences, and Vimalan Biosciences; and that his spouse has served as a consultant for and owns stock options in Ophotech and Progenity; is an employee of and owns stock options in Oppilan Pharma, Escalier Biosciences; Precision IBD (also owns Prometheus Laboratories), Ventyx Biosciences, and Vimalan Biosciences. Lisa M. Shackleton reports consulting fees from Roberts Clinical Trials, Inc. Larry W. Stitt and Claire E. Parker are employees of Roberts Clinical Trials, Inc. Peter Bossuyt reports educational grants from AbbVie, Mundipharma, Pfizer, Janssen, and Mylan; speaker's fees from AbbVie, Takeda, Pfizer, and Janssen; and advisory board fees from Hospira, Janssen, MSD, Mundipharma, Roche, Pfizer, Sandoz, Takeda, AbbVie, Mylan, and Pentax. Mark Löwenberg reports having served as a speaker and/or principal investigator for AbbVie, Covidien, Dr Falk, Ferring Pharmaceuticals, Merck Sharp & Dohme, Receptos, Takeda, and Tramedico; and research grants from AbbVie, Merck Sharp & Dohme, and Achmea Healthcare. Reena Khanna reports consulting fees from AbbVie, Janssen, Merck, Takeda, Roche/Genetec AbbVie, Janssen, Merck, Takeda, and Roche/Genetec; having served on an advisory board for AbbVie, Janssen, Merck, Takeda, Roche/Genetec, and Pfizer; and speaker's fees from AbbVie, Janssen, Takeda, Shire, and Pendo-pharm. Severine Vermeire reports grants/research support from MSD, AbbVie, Takeda, Janssen, and Pfizer; honoraria or consultation fees from AbbVie, MSD, Takeda, Ferring, Genentech/Roche, Shire, Pfizer Inc, Galapagos, Roberts Clinical Trials, Inc, Mundipharma, Hospira, Celgene, Second Genome, Progenity, GSK, Lilly, Arena, Gilead, and Janssen; and participation in a company-sponsored speaker's bureau for AbbVie, MSD, Takeda, Ferring, Hospira, Pfizer,

Janssen, and Tillots. Paul Rutgeerts reports consulting fees from Johnson & Johnson, MSD, AbbVie, Takeda, BMS, Pfizer, Genentech, Tillotts Pharma, Celgene, and Roberts Clinical Trials, Inc; having served as a Scientific Advisory Board member for Johnson & Johnson, MSD, AbbVie, Takeda, BMS, Pfizer, Genentech, and Tillotts Pharma; payments for lectures/speakers bureau from Johnson & Johnson, MSD, AbbVie, and Takeda; and research grants from Johnson & Johnson, MSD, and AbbVie. Brian G. Feagan reports fees for serving on advisory boards from AbbVie, Allergan, Amgen, AstraZeneca, Avaxia Biologics, Boehringer Ingelheim, Bristol-Myers Squibb, Celgene, Elan-Biogen, Ferring, Genentech-Roche, Janssen-Johnson & Johnson, Merck, Millennium, Nestlé, Novo Nordisk, Novartis, Pfizer, Prometheus, Protagonist, Receptos, Salix, Sigmoid Pharma, Takeda, Teva, TiGenix, Tillotts Pharma, and UCB Pharma; consulting fees from AbbVie, Actogenix, Akros, Albireo Pharma, Allergan, Amgen, AstraZeneca, Avaxia Biologics, Avir Pharma, Axcan, Baxter Healthcare, Biogen Idec, Boehringer Ingelheim, Bristol-Myers Squibb, Calypso Biotech, Celgene, Elan-Biogen, EnGene, Ferring, Genentech-Roche, Glaxo Pharma, Gilead Sciences, Given Imaging, GlaxoSmithKline, Ironwood, Janssen Biotech-Centocor, Janssen-Johnson & Johnson, Kyowa Hakko Kirin, Eli Lilly, Merck, Mesoblast Pharma, Millennium, Nestlé, Novo Nordisk, Novartis, Pfizer, Prometheus, Protagonist, Receptos Salix, Sanofi, Shire, Sigmoid Pharma, Synergy Pharma, Takeda, Teva, TiGenix, Tillotts Pharma, UCB Pharma, Vertex, VHSquared, Wyeth, Zealand, and Zygenia; lecture fees from AbbVie, Janssen-Johnson & Johnson, Takeda, and UCB Pharma; and grant support from AbbVie, Amgen, AstraZeneca, Bristol-Myers Squibb, Janssen Biotech-Centocor, Janssen-Johnson & Johnson, Pfizer, Receptos, Sanofi, and Takeda; and being a consultant to Roberts Clinical Trials, Inc. Vipul Jairath reports consulting fees from AbbVie, Eli Lilly, GlaxoSmithKline, Arena Pharmaceuticals, Genetech, Pendopharm, Sandoz, Merck, Takeda, Janssen, Roberts Clinical Trials, Inc, Topivert, and Celltrion; and speaker's fees from Takeda, Janssen, Shire, Ferring, AbbVie, and Pfizer. Geert R. D'Haens reports having served as an advisor for AbbVie, Ablynx, Amakem, AM Pharma, Avaxia, Biogen, Bristol-Myers Squibb, Boehringer Ingelheim, Celgene, Celltrion, Cosmo, Covidien, Ferring, Dr Falk Pharma, Engene, Galapagos, Gilead, GlaxoSmithKline, Hospira, Immunic, Johnson and Johnson, Lycera, Medicmetrics, Millennium/Takeda, Mitsubishi Pharma, Merck Sharp & Dohme, Mundipharma, Novo Nordisk, Pfizer, Prometheus Laboratories/Nestlé, Protagonist, Receptos, Roberts Clinical Trials, Inc, Salix, Sandoz, Setpoint, Shire, Teva, Tigenix, Tillotts, Topivert, Versant, and Vifor; and speaker's fees from AbbVie, Ferring, Johnson & Johnson, Merck Sharp & Dohme, Mundipharma, Norgine, Pfizer, Shire, Millennium/Takeda, Tillotts, and Vifor. The remaining author discloses no conflicts.

Q8

Supplemental Methods

This was a prospective multicenter study conducted at Amsterdam UMC, University Medical Center (Amsterdam, the Netherlands), and Imelda General Hospital (Bonheiden, Belgium).

Videos from 50 patients were de-identified and independently and randomly assessed in triplicate by four experienced central reader gastroenterologists blinded to all clinical information. Video recordings included examination of the colon, anastomosis, and neoterminal ileum. To accurately evaluate intrarater reliability, there was a 2-week interval between evaluations for the same patient to achieve memory extinction.

All data were collected in a standardized electronic case report form. Patient demographic and disease-related data were extracted from electronic medical records. Mucosal appearance was evaluated using:

1) Rutgeerts score (RS)

- i0: no lesions
- i1: ≤ 5 aphthous lesions
- i2: > 5 aphthous lesions with normal mucosa between the lesions, or skip areas of larger lesions or lesions confined to the ileocolonic anastomosis
- i3: Diffuse aphthous ileitis with diffusely inflamed mucosa
- i4: Diffuse inflammation with already larger ulcers, nodules, and/or narrowing

2) Modified Endoscopic Postoperative Recurrence Score (mEPRS)

- i0: No lesions in the distal ileum
- i1: ≤ 5 aphthous lesions in the distal ileum
- i2a: Lesions confined to the ileocolonic anastomosis (including anastomotic stenosis)
- i2b: >5 aphthous ulcers or larger lesions, with normal mucosa in between, in the neoterminal ileum (with or without anastomotic lesions)

- i3: Diffuse aphthous ileitis with diffusely inflamed mucosa
- i4: Large ulcers with diffuse mucosal inflammation or nodules or stenosis in the neoterminal ileum

3) A global evaluation of lesion severity (GELS): the GELS is a 10-cm linear analog scale ranging from 0 (no endoscopic disease) to 10 (worst disease activity ever seen).

Inter- and intrarater reliability was quantified by the intraclass correlation coefficient (ICC), which is equivalent to the weighted kappa in the case of ordinal data. Associated 2-sided 95% confidence intervals were obtained using the cluster bootstrap method (nonparametric centile bootstrapping) with 2000 samples obtained with replacement at the level of the video to maintain data structure. The strength of agreement was interpreted according to the criteria established by Landis and Koch¹ as "slight" (ICC 0.00–0.20), "fair" (ICC 0.21–0.40), "moderate" (ICC 0.41–0.60), "substantial" (ICC 0.61–0.80), or "almost perfect" (ICC 0.81–1.00). The validity of the RS and mEPRS was evaluated by correlating these scores to the GELS and estimated using mixed model methods to account for repeated ratings by each central reader.

We calculated the degree of misclassification that would result in inappropriate therapeutic initiation based on three treatment thresholds ($\geq i2a$, $\geq i2b$, and $> i2b$), assuming treatment is required for patients with RS $> i1$ and that if three readers agreed on the mEPRS relative to the threshold for treatment initiation, selecting the fourth reader who disagreed would have resulted in misclassification. Under a binomial distribution, we calculated the probability that a single evaluation would lead to inappropriate treatment based on the number of cases in which disagreement was observed when one of four reader's assessment was randomly selected.

All analyses were conducted using SAS version 9.4 (SAS Institute Inc, Cary, NC).

References

1. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.

Supplemental Table 1. Patient Characteristics and Endoscopic Findings

Patient Characteristics	n=50
Male	18 (36)
Age, y	43.2 (19–73)
Disease duration before ileocolic resection, y	5.7 (0–23)
Montreal classification ^a	
Age at diagnosis	
≤16 y	7 (14)
17–20 y	33 (66)
>40 y	5 (10)
Missing age of diagnosis	5 (10)
Disease location	
L1 (ileal)	25 (50)
L2 (colonic)	0 (0)
L3 (ileocolic)	24 (48)
Missing disease location	1 (2)
Disease behavior	
B1 (stricturing)	18 (36)
B2 (penetrating)	19 (38)
B3 (non-stricturing, non-penetrating)	12 (24)
Perianal disease	16 (33)
Missing disease behavior	1 (2)
Medications at time of colonoscopy	
No medical therapy	24 (48)
5-aminosalicylate	3 (6)
Budesonide	2 (4)
Thiopurines (6-mercaptopurine or azathioprine)	5 (10)
Methotrexate	1 (2)
Biologic	9 (18)
Combination therapy	4 (8)
Missing medication data	2 (4)
Distribution of endoscopic lesions based on all video reads (n = 599) ^a	
i0	125 (21)
i1	49 (8)
i2a	103 (17)
i2b	136 (23)
i3	46 (8)
i4	140 (23)

Values are n (%) or mean (range).

^aVideos from 50 patients were assessed by four readers three times (600 independent assessments, one missing assessment)