Journal of Crohn's and Colitis, 2016, 1001–1005
doi:10.1093/ecco-jcc/jjw082
Advance Access publication April 11, 2016
Original Article

Original Article

Interobserver Variation Study of the Rutgeerts Score to Assess Endoscopic Recurrence after Surgery for Crohn's Disease

Philippe Marteau,[a,b] David Laharie,[c] Jean-Frédéric Colombel,[d,e] Laurence Martin,[a] Hugues Coevoet,[d] Matthieu Allez,[f] Guillaume Cadiot,[g] Arnaud Bourreille,[h] Franck Carbonnel,[i] Yoram Bouhnik,[j] Benoit Coffin,[k] Bernard Duclos,[l] Jean Louis Dupas,[m] Jacques Moreau,[n] Edouard Louis,[o] Jean-Yves Mary[p] on behalf of the GETAID

[a]Service d'Hépatogastroentérologie, Hôpital Saint Antoine, 184 rue du Faubourg Saint Antoine, 75012 Paris cedex, France [b]Denis Diderot – Paris7 University, Paris, France [c]CHU de Bordeaux, Hôpital Haut-Lévêque, Service d'Hépato-gastroentérologie – Université Bordeaux, Laboratoire de bactériologie, F-33000 Bordeaux, Pessac, France [d]CHRU de Lille, Hôpital Claude Huriez, Service des Maladies de l'Appareil Digestif -Endoscopie Digestive, Lille, France [e]Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA [f]Department of Hepatogastroenterology, Hôpital Saint-Louis, Paris, France [g]Department of Hepato-Gastroenterology and Digestive Oncology, Hôpital Robert Debré, Boulevard du Général Koenig, 51100 Reims Cedex, France [h]CHU de Nantes, Hôtel-Dieu, Hépato-Gastroentérologie, Institut des Maladies de l'Appareil Digestif, F-44093 Nantes, France [i]Department of Gastroenterology, Assistance Publique-Hôpitaux de Paris (AP-HP), University Hospitals Paris-Sud, Site de Bicêtre, Paris Sud University, Paris XI, Le Kremlin Bicêtre, Villejuif, France [j]Hôpital Beaujon, Gastroentérologie, Maladies Inflammatoires Chroniques de l'Intestin et Assistance Nutritive, APHP— Université Paris Diderot Paris 7, Clichy, France [k]Hôpital Louis Mourier, service d'Hépato-Gastroentérologie, Pôle Maladie Appareil Digestif, APHP – Université Paris VII, F-92700 Colombes, France [l]Service d'Hépato-Gastroentérologie et Assistance Nutritive, CHU Strasbourg, Strasbourg, France [m]Service d'Hépato- Gastroentérologie, CHU Amiens, Université de Picardie Jules Verne, Amiens, France [n]CHU de Toulouse, Hôpital Rangueil, Service de Gastro-entérologie et Nutrition, F-31059 Toulouse, France [o]Centre Hospitalier, Universitaire de Liège, Liège, Belgium [p]Inserm UMR 1153, Equipe Epidemiologie Clinique, Statistique pour la Recherche en Santé, Hôpital Saint-Louis, Université Paris Diderot – Paris 7, Paris, France
GETAID: Groupe d'Etudes Thérapeutiques des Affections Inflammatoires du Tube Digestif, St Louis Hospital, Paris France

Corresponding author: David Laharie, MD, PhD, service d'Hépato-gastroentérologie, Hôpital Haut-Lévêque, CHU de Bordeaux, 33600 Pessac, France. Tel: +33-557-656-439; E-mail: david.laharie@chu-bordeaux.fr

Abstract

Background: After resection surgery for Crohn's disease, recurrence of endoscopic lesions at the site of the anastomosis or in the neoterminal ileum is graded according to the Rutgeerts score (RS). The goal of this study was to test the interobserver variability for RS.

Methods: Thirteen trained endoscopists evaluated the RS on 39 videotapes of patients who had undergone resection for Crohn's disease with an ileocolonic anastomosis 6 months earlier. Videotapes were randomly assigned to endoscopists through a balanced incomplete block design. Each videotape was scored independently by four endoscopists, and each endoscopist evaluated 12 videotapes, making a total of 156 videotape assessments. Reproducibility levels of the RS were assessed through unweighted kappa estimates among multiple raters. The proportion of inappropriate therapeutic initiation was estimated by randomly selecting one endoscopist for each videorecording, assuming that the majority of endoscopists correctly classified endoscopic recurrence.

Results: The kappa estimates were 0.43 (95% confidence interval: 0.33–0.52) for the RS on a 5-grade scale, 0.47 (0.28–0.66) for RS < i2 vs. ≥ i2, and 0.64 (0.42–0.85) for RS ≤ i2 vs. > i2. The percentages of inappropriate therapeutic initiation were 12.8% (3.8–21.9) when initiation was triggered by a RS ≥ i2 and 8.3% (1.1–15.6) when initiation was triggered by a RS > i2 ($p = 0.41$).

Conclusion: The reproducibility of the RS was moderate, especially when differentiating <i2 from ≥i2, which may lead to incorrect therapeutic decisions in >10% of patients.

Key Words: Crohn's disease; Endoscopy; postoperative endoscopic recurrence; reproducibility; Rutgeerts score

Introduction

Despite recent advances in medical management, many patients with Crohn's disease (CD) still require intestinal surgery throughout their lives.[1] Surgery is not a cure, and postoperative recurrence is common in patients with CD.[2] Ileocolonoscopy is considered to be the gold standard in the diagnosis and monitoring of CD endoscopic recurrence after surgery, both in clinical trials and clinical practice.[3–11] Endoscopic activity at the anastomotic site and in the neoterminal ileum after ileocolonic resection is graded using a scoring system developed by Rutgeerts *et al*.[12] Endoscopic recurrence is commonly defined as a score of i2–i4 (i.e. ≥i2) and used to initiate a new therapy in an attempt to prevent clinical recurrence in the following months. However, since its description in 1990, the properties of the Rutgeerts score (RS) have been poorly studied.

The aim of the present interobserver variation study was to assess the reproducibility level of the RS and to assess how this could affect therapeutic decision-making after surgery.

Patients and methods

The present study was an interobserver variation study performed on videorecorded ileocolonoscopies of patients recruited and followed-up in either of two clinical trials conducted in 11 French centres from the Groupe d'Etude Thérapeutique des Affections Inflammatoires du tube Digestif (GETAID). These two controlled trials (GETAID 2001-1 and GETAID 2001–2) evaluated CD postoperative recurrence in patients receiving 5-aminosalicylate acids (5-ASA), azathioprine (unpublished data), *Lactobacillus johnsonii* La1, or placebo, respectively.[11] Inclusion criteria for the present study were the following: adults with CD diagnosed according to the usual criteria, bowel resection (right ileocolectomy or colectomy with ileocolic or ileorectal anastomosis) removing all macroscopic intestinal lesions, anastomosis accessible by standard colonoscopy, and complete endoscopic evaluation 6 months after surgery and with videorecording. Patients with colo-colic or colorectal anastomosis and patients whose ileocolonoscopy was not recorded were excluded. All subjects gave written informed consent for these studies, which were approved by the Ethics Committee.

Endoscopists involved in the interobserver variation study were operators trained in inflammatory bowel disease and using RS in current practice. Allocation of the videos to be assessed by each of the 13 endoscopists was performed through a balanced incomplete block randomization so that each of the 39 videos was evaluated independently by four endoscopists blinded to the patient's characteristics. Each endoscopist evaluated 12 videos, and 156 videotape assessments were performed (detailed data are provided in the supplementary material table). Endoscopists evaluated the RS using its definition and five-grade scale (i0 to i4) (Table 1).[12]

The reproducibility levels of the RS were assessed through unweighted kappa estimates among multiple raters.[13] Agreement was determined for the RS relative to the cut-off of ≥i2 and relative to the cut-off of >i2 being used to define severe endoscopic recurrence). The agreement was considered as 'moderate' for kappa values of 0.41–0.60, 'substantial' for kappa values of 0.61–0.80, and almost perfect for kappa values above 0.80.[14]

We first assumed that a score of ≥i2 would trigger initiation of medical therapy. This cut-off value has been used as the main end point in most clinical trials studying the prevention of postoperative recurrence in CD.[3,6–11] When using this assumption, the four assessors may disagree on the need for medical therapy because they disagree on RS relative to i2; for instance, one may score i1 and the three others i2. To estimate the proportion of inappropriate therapeutic initiations, we assumed that when three assessors agreed on the RS relative to the cut-off used to initiate intervention (i.e. three evaluations under the cut-off or, alternatively, three evaluations at or above the cut-off), their evaluation led to an appropriate therapeutic initiation. Thus, the evaluation of the last assessor, the one who disagreed with them, would have led to an inappropriate therapeutic initiation. Consequently, if we selected at random one of the four assessors when three assessors agreed and the fourth differed, 25% of the selections would have led to an inappropriate therapeutic initiation. When two assessors made an evaluation indicating therapeutic initiation, whereas the other two made an evaluation indicating the opposite decision, any assessor had a 50% chance of recommending appropriate therapeutic initiation according to the criteria. Consequently, if we selected at random one of the four assessors when two assessors agreed on therapeutic initiation and two on no therapeutic initiation, half the selections would have led to an inappropriate therapeutic initiation. In all other situations, all assessors agreed, and if we selected at random one assessor, none would have indicated an inappropriate therapeutic initiation. According to the number of cases in which each situation was observed, it was possible to calculate, through binomial distribution, the probability that an evaluation would have led to an inappropriate therapeutic initiation when selecting at random one endoscopist out of the four who assessed each of the 39 videorecordings (see the supplementary appendix for detailed calculations).

Table 1: The Rutgeerts score of ileal and anastomotic lesions.[12]

i0: no lesion
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 larger ulcers, nodules, and/or narrowing

The same type of calculation was repeated, but assuming that a score of >i2 was used to initiate a new treatment.

These calculations allowed us to estimate the difference in the proportions of inappropriate therapeutic initiations (with the 95% confidence interval) when using ≥i2 versus ≥i3 as the cut-off RS, taking into account the pairing of these data.

All results (kappa values and the probabilities of an evaluation leading to an appropriate therapeutic initiation) are presented as estimate ± standard error (SE) (95% confidence interval). Analyses were performed with the software SPSS®.

Results

The clinical characteristics of the 39 patients are displayed in Table 2.

The unweighted kappa ± SE estimate for the RS was 0.43 ± 0.05 (0.33–0.52) on a 5-grade scale, showing 'moderate' agreement. The kappa estimate ± SE was 0.47 ± 0.10 (0.28–0.65) for the RS relative to the cut-off value of i2 or more ('moderate' agreement), and 0.64 ± 0.11 (0.42–0.85) for the RS relative to the cut-off value of >i2 ('substantial' agreement).

When initiating treatment for a RS of ≥i2, assuming that the majority of assessors made the correct evaluation for appropriate therapeutic initiation, the estimate of the number of evaluations leading to an inappropriate therapeutic initiation was 5.00, with a variance of 3.25, leading to an estimate of the probability of inappropriate therapeutic initiation of 12.8% (3.8–21.9), which corresponds to one in eight patients. When deciding on treatment for a RS of >i2, the estimated number of evaluations leading to an inappropriate therapeutic initiation was 3.25, with a variance of 2.06, leading to an estimate of the probability of inappropriate therapeutic initiation of 8.3% (1.1–15.6). The estimated difference between the number of evaluations leading to an inappropriate therapeutic initiation according to the cut-off of ≥i2 versus >i2 was 1.75, with a variance of 4.44, taking into account the pairing of these two numbers evaluated on the same videotapes, leading to a difference in the probability of inappropriate therapeutic initiation of 4.5% (−6.1–15.1) (p = 0.41).

Table 2: Characteristics of the 39 patients.

Variable	*n* (%) or median (interquartile range)
Age (years)	32 (28–38)
Male gender	14 (36)
Smokers	11 (28)
Time elapsed since diagnosis of CD (months)	41 (17–102)
Previous surgery for CD	9 (23)
Disease location (*n* = 38)	
small bowel	24 (63)
small bowel + colon	13 (34)
colon	1 (3)
Disease behaviour	
non-stenosing, non-penetrating	0 (0)
stenosing only	22 (57)
Penetrating	13 (33)
NA	4 (10)
Months between surgery and ileocolonoscopy	6.1 (5.9–6.4)
Rutgeerts score (156 assessments, 4 per video)	
i0	24 (15)
i1	22 (14)
i2	69 (44)
i3	15 (10)
i4	26 (17)

NA, not available.

Discussion

To our knowledge, the present study is one of the first that has formally assessed the interobserver variation for RS. In addition, it attempted to derive the consequences of the variation in RS evaluation in terms of a patient's management. The main result is that the reproducibility of RS evaluation on a 5-grade scale was moderate, as was its evaluation relative to a cut-off of < i2 versus a cut-off > or = i2 indicating therapeutic initiation. As a consequence of the latter variation, we estimated that one therapeutic decision out of eight could be inappropriate.

Only three other studies have estimated the reproducibility level of the RS score on a 5-grade scale. In the Daperno *et al.* study, 10 postoperative CD colonoscopies were evaluated by 14 experts, leading to 140 assessments.[15] In the Kennedy *et al.* study, 43 colonoscopies were evaluated by five experts, leading to 215 assessments.[16] In the Gesce *et al.* study, 25 colonoscopies were evaluated three times by four experts, leading to 100 independent assessments.[17] In our study, 13 experts were involved in the evaluation of 39 colonoscopies, leading to 156 assessments. As a consequence of the assignment of assessments through an incomplete balanced block design, the endoscopist and colonoscopy samplings were rather large in our study, with the total number of assessments being similar to that in the three previous studies.

The kappa estimates for RS evaluation on a 5-grade scale were quite similar in three studies: 0.53 in the Daperno *et al.* study,[15] 0.50 in the Kennedy *et al.* study,[16] and 0.43 in the present study, confirming a moderate agreement. In the Gesce *et al.* study, the intraclass correlation coefficient (ICC) was 0.72,[17] a level quite similar to those obtained with the same statistics in the Kennedy *et al.* study (0.82)[16] and in the present study (0.76). Nevertheless, when using ICC instead of unweighted kappa to estimate the agreement level of the RS on a 5-grade scale, much higher levels of agreement were systematically obtained, corresponding to an almost perfect agreement when applying the same rules to the ICC as those used for kappa. This apparent contradiction is the consequence of the equivalence between ICC and weighted kappa for ordinal data that are using unit distance between adjacent categories and square distance weights for weighted kappa. Weighted kappa consists of considering what the disagreements between observers' quotations contribute partially to agreement if they are not extreme. This contribution is quantified by a decreasing weight when quotations diverge on the ordinal scale. When using the square distance weighted kappa for RS, weights would be: 1 for perfect agreement, 0 for maximal disagreement (i0 and i4), 0.75 in the case of disagreement between i0 and i2, i1 and i3, or i2 and i4, and 0.44 in the case of disagreement between i0 and i3, or i1 and i4. It means that two disagreements from i0 and i3, having a total weight of 0.88, would provide a similar contribution to agreement than one perfect agreement weighted 1.00. This demonstrates that, converse to unweighted kappa, ICC overestimates agreement level of RS on a 5-grade scale, by giving too much weight to major disagreements. This criticism holds for recent publications using widely ICC to assess agreement level of ordinal scores in inflammatory bowel disease (see for instance Mosli *et al.*[18] for histological assessment of disease severity in ulcerative colitis).

The kappa estimates for RS evaluation relative to <i2 versus ≥i2 were 0.83 in the Kennedy *et al.* study[16] and 0.47 in our study. This large difference could be the consequence of colonoscopy sampling. Indeed, in the Kennedy *et al.* study,[16] only seven colonoscopies (16%) were rated i2 and 15 (35%) i3 or i4 by the original endoscopist, compared with 19 (49%) and 4 (10%) in our study,

respectively. It is likely that agreement would be higher when dealing with numerous grade 3 and 4 colonoscopies than when dealing with numerous grade 2 colonoscopies. Differences in RS distributions in the two samples could explain the variation in kappa estimates. Moreover, as described in the seminal paper from Rutgeerts *et al.* and in a recent review from the International Organisation of Inflammatory Bowel Disease (IOIBD), the i2 subscore includes various postoperative mild lesions on the anastomosis and/or terminal ileum of great heterogeneity and of intermediate prognostic value.[19]

Data from the present study are of importance when considering clinical trials that have looked at the efficacy of various drugs to prevent CD postoperative recurrence using the i2 threshold.[3–11] Meta-analysis of studies testing placebo, 5-ASA, thiopurines, or anti-TNF agents in this setting have shown the superiority of biologics over all other treatments.[20] However, the interobserver variation observed in the present study may lead to an erroneous assessment with potential therapeutic consequences in at least 13% of the cases. When considering only severe postoperative recurrence, defined as a RS >i2, this meta-analysis suggests that biologics were more effective than immunosuppressants, with a difference not reaching statistical significance.[20] It should be noted that, for this threshold, the proportion of erroneous therapeutic decisions was only 8% in the present series. This proportion was not significantly different from the 13% estimated for a RS of ≥i2 or more, probably due to the small number of colonoscopies graded i3 or i4 in our sample. In clinical practice, it does not mean that therapeutic decisions would be more appropriate when using an i3 threshold instead of an i2 threshold: treating only patients with RS >i2 would probably avoid overtreatment but may undertreat some patients with a high risk of clinical relapse. Development of new tools for assessing CD postoperative recurrence in addition to RS, such as magnetic resonance imaging or faecal calprotectin, may help the physician to refine the best therapeutic approach in this situation.

One of the strengths of the present study was the experience of and the large number of endoscopists who had previously used the RS for years in daily practice and clinical trials.[11] We acknowledge, that despite four-fold assessments, the low number of patients with severe endoscopic lesions is a limitation. Furthermore, it has not been possible to assess in the same time the quality of the videos, especially bowel preparation and recording.

In conclusion, the variability of the RS appears large and the reproducibility for lesions i2 or more was only moderate. Thus, when solely based on the RS, management of patients with CD postoperative recurrence may lead to wrong therapeutic initiations. Reasons for the variation in the RS evaluation should be studied in an attempt to increase its reproducibility.

Funding

None.

Conflict of Interest

None.

Acknowledgments

Thanks to M. Lémann who initiated and designed the study, participated in patient recruitment and the ileocolonoscopy assessments.

L. Martin and H. Coevoet defended their MD thesis using some of the data presented in this paper.

Author Contributions

PM, DL, JFC: study conception and design, investigation (assessing videorecordings), data interpretation, drafting and critical revision of the manuscript.

LM, HC: data collection.

MA, GC, AB, FC, YB, BC, BD, JLD, JM, EL: investigation (assessing videorecordings).

JYM: statistical analysis, drafting and critical revision of the manuscript.

Supplementary Data

Supplementary data to this article can be found online at *ECCO-JCC* online.

References

1. Dignass A, Van Assche G, Lindsay JO, *et al*. European Crohn's and Colitis Organisation (ECCO). The second European evidence-based Consensus on the diagnosis and management of Crohn's disease: current management. *J Crohns Colitis* 2010;**4**:28–62.
2. Buisson A, Chevaux JB, Allen PB, *et al*. Review article: the natural history of postoperative Crohn's disease recurrence. *Aliment Pharmacol Ther* 2012;**35**:625–33.
3. De Cruz P, Kamm MA, Hamilton AL, *et al*. Crohn's disease management after intestinal resection: a randomised trial. *Lancet* 2014;**385**:1406–17.
4. Wright EK, Kamm MA, De Cruz P, *et al*. Measurement of fecal calprotectin improves monitoring and detection of recurrence of Crohn's disease following surgery. *Gastroenterology* 2015;**148**:938–47.
5. Yamamoto T, Bamba T, Umegae S, Matsumoto K. The impact of early endoscopic lesions on the clinical course of patients following ileocolonic resection for Crohn's disease: a 5-year prospective cohort study. *United European Gastroenterol J* 2013;**1**:294–8.
6. Mañosa M, Cabré E, Bernal I, *et al*. Addition of metronidazole to azathioprine for the prevention of postoperative recurrence of Crohn's disease: a randomized, double-blind, placebo-controlled trial. *Inflamm Bowel Dis* 2013;**19**:1889–95.
7. Herfarth HH, Katz JA, Hanauer SB, *et al*. Ciprofloxacin for the prevention of postoperative recurrence in patients with Crohn's disease: a randomized, double-blind, placebo-controlled pilot study. *Inflamm Bowel Dis* 2013;**19**:1073–9.
8. Reinisch W, Angelberger S, Petritsch W, *et al*.; International AZT-2 Study Group. Azathioprine versus mesalazine for prevention of postoperative clinical recurrence in patients with Crohn's disease with endoscopic recurrence: efficacy and safety results of a randomised, double-blind, double-dummy, multicentre trial. *Gut* 2010;**59**:752–9.
9. Regueiro M, Schraut W, Baidoo L, *et al*. Infliximab prevents Crohn's disease recurrence after ileal resection. *Gastroenterology* 2009;**136**:441–50.
10. D'Haens GR, Vermeire S, Van Assche G, *et al*. Therapy of metronidazole with azathioprine to prevent postoperative recurrence of Crohn's disease: a controlled randomized trial. *Gastroenterology* 2008;**135**:1123–9.
11. Marteau P, Lémann M, Seksik P, *et al*. Ineffectiveness of *Lactobacillus johnsonii* LA1 for prophylaxis of postoperative recurrence in Crohn's disease: a randomised, double blind, placebo controlled GETAID trial. *Gut* 2006;**55**:842–7.
12. Rutgeerts P, Geboes K, Vantrappen G, *et al*. Predictability of the postoperative course of Crohn's disease. *Gastroenterology* 1990;**99**:956–63.
13. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**:378–82.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
15. 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**:969–73.
16. Kennedy NA, Ennis H, Gaya DR *et al*. Interobserver agreement in assessment of Rutgeerts score of endoscopic recurrence of ileal Crohn's disease. *Gastroenterology* 2015;**64**:A243.

17. Gesce KB, Lowenberg M, Bossuyt P, *et al*. Agreement among experts in the endoscopic evaluation of postoperative recurrence in Crohn's disease using the Rutgeerts score. *Gastroenterology* 2014;**146**:S-227.
18. Mosli MH, Feagan BG, Zou G, *et al*. Reproducibility of histologic assessment of disease activity in UC. *Gut* 2015;**64**:1765–73.
19. Vuitton L, Marteau P, Sandborn WJ, *et al*. IOIBD technical review on endoscopic indices for Crohn's disease clinical trials. *Gut* 2015, published online Sep 9 doi: 10.1136/gutjnl-2015-309903.
20. Yang Z, Ye X, Wu Q, Wu K, Fan D. A network meta-analysis on the efficacy of 5-aminosalicylates, immunomodulators and biologics for the prevention of postoperative recurrence in Crohn's disease. *Int J Surgery* 2014;**12**:516–22.