Original Contribution Kitasato Med J 2021; 51: 1-9

Comparison of the reproducibility of endoscopic scores in patients with ulcerative colitis: MES, UCEIS, and EAI scores

Kana Kawagishi,[1] Kaoru Yokoyama,[1] Kiyonori Kobayashi,[2] Wasaburo Koizumi[1]

[1] Department of Gastroenterology, Kitasato University School of Medicine
[2] Research and Development Center for New Medical Frontiers, Kitasato University School of Medicine

Background: Endoscopic scores to assess the severity of ulcerative colitis (UC) should be simple and reproducible.
Objective: Three types of endoscopic scores, the Mayo endoscopic subscore (MES), the Ulcerative Colitis Endoscopic Index of Severity (UCEIS), and the Endoscopic Activity Index (EAI), were used to examine the level of agreement among raters and the consistency for each rater and to obtain highly reproducible endoscopic scores.
Methods: Using 20 sheets of endoscopic images of UC, 6 experts and 20 trainees of gastrointestinal endoscopy assessed the severity of intestinal lesions according to the MES, UCEIS, and EAI. The level of agreement among raters for endoscopic scores was assessed by using Krippendorff's α (*alpha*) coefficients. We compared not only the α coefficients for each endoscopic score, but also the level of agreement according to the examiners' endoscopic experience and the variables of each score to determine intraobserver consistency. Among the experts, we compared Krippendorff's α coefficients for each score for consistency with each of the raters, i.e., interobserver consistency.
Results: The Krippendorff's α coefficient in all the raters was 0.808 for MES scores, 0.840 for UCEIS scores, and 0.866 for EAI scores. The α coefficient of the EAI score was highest, with a significant difference from those of MES and UCEIS scores. On comparing alpha coefficients for each score according to the examiners' endoscopic experience, no difference was found in the α coefficient by experts. In trainees, however, the α coefficient was significantly higher in the EAI than that in the MES. No difference was found in the α coefficient, interobserver consistency for each rater.
Conclusions: Among endoscopic scores of UC, the EAI was associated with the highest level of agreement among the raters and were not influenced by their endoscopic experience. However, because the EAI is based on many variables, the development of a highly reproducible and convenient endoscopic scoring system is required.

Key words: ulcerative colitis, endoscopic scores, level of agreement, consistency

Introduction

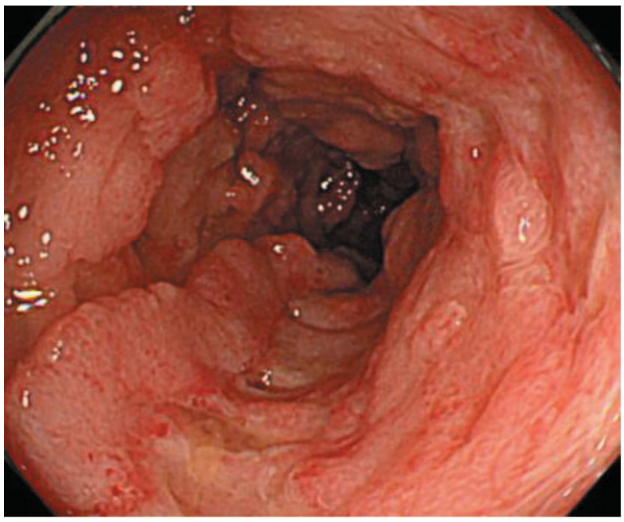
To assess the severity of active ulcerative colitis (UC), evaluation by colonoscopy is essential in addition to an assessment based on clinical symptoms and blood tests. Colonoscopic findings also play an important role in the evaluation of treatment response. An improvement in clinical symptoms and mucosal healing are a recent treatment goal for patients with UC.[1] To date, several different types of endoscopic scores have been used to evaluate the severity of UC. However, endoscopic scores are not consistant among hospitals and clinical studies.

Endoscopic scores used to evaluate the severity of UC are expected to be able to reflect clinical symptoms, treatment response, and outcomes and to be highly reproducible and convenient among raters and consistent for each rater, i.e., intra- and interobserver consistency. The Mayo endoscopic subscore (MES), which consists of endoscopic variables extracted from Mayo scores[2] is a convenient score that can comprehensively evaluate parameters, such as vascular pattern, mucosal redness, bleeding, erosion, and ulceration, and has been widely used to evaluate the endoscopic severity in trials of new drugs for UC and clinical studies.[2-4] However, the MES

Received 19 August 2020, accepted 5 October 2020
Correspondence to: Kana Kawagishi, Department of Gastroenterology, Kitasato University School of Medicine
1-15-1 Kitasato, Minami-ku, Sagamihara, Kanagawa 252-0374, Japan
E-mail: k-kana@kitasato-u.ac.jp

has been reported as not having a high level of agreement among raters.[5,6] The Ulcerative Colitis Endoscopic Index of Severity (UCEIS),[7] which has recently been advocated, is a scoring system proposed to evaluate severity on the basis of the sum of 3 variables, vascular pattern, bleeding, erosion and ulcers, has been reported to have a high level of agreement among raters.[7] In Japan, Naganuma et al.[8] advocated the Endoscopic Activity Index (EAI). This score consists of 6 variables, such as the size and depth of ulcers, bleeding, mucosal edema, redness, and mucous exudate, for evaluation.

The MES can be used to easily evaluate the severity of colonoscopic findings. However, when evaluating the response of active UC to various treatments, the score does not change much and cannot reflect treatment response if active lesions partially remain even though the endoscopic findings are generally improved. To accurately evaluate the short-term response of intestinal lesions to various treatments in patients with UC, we believe that endoscopic scores with different evaluation methods than those from the MES should be used.[9] Some studies have examined the reproducibility of individual endoscopic scores intra- and interobserver raters.[5-7] To our knowledge, however, no studies have compared the reproducibility of several endoscopic scores. We therefore compared the level of agreement intra- and interobserver raters for 3 types of endoscopic scores, MES, UCEIS, and EAI, to obtain highly reproducible endoscopic scores.



		Points				Total
MES	—	0	1	2	(3)	3
UCEIS	Vascular pattern	0	1	(2)	—	5
	Bleeding	(0)	1	2	3	
	Erosion and ulcers	0	1	2	(3)	
EAI	Size of ulcers	0	1	(2)	3	7
	Depth of ulcers	0	1	2	(3)	
	Bleeding	(0)	1	2	3	
	Mucosal edema	0	(1)	2	3	
	Redness	0	(1)	2	—	
	Mucous exudate	(0)	1	2	—	

MES, Mayo endoscopic subscore; UCEIS, ulcerative colitis endoscopic index of severity; EAI, endoscopic activity index

Figure 1. Endoscopic image and endoscopic scores in a patient with ulcerative colitis

Methods

In 20 patients with UC who underwent colonoscopy in Kitasato University Hospital or Kitasato University East Hospital from April 2013 through Feburary 2015, a total of 20 sheets of white-light endoscopic images were evaluated. Among intestinal lesions of patients with UC, active inflammation was observed in 16 sheets but not in 4 sheets. Regarding the selection of the images to be evaluated, different endoscopic images of intestinal lesions with mild to severe inflammation were selected on the basis of endoscopic findings, such as mucosal redness and edema, ulceration, erosion, and friability. Endoscopic images were selected arbitrarily by 2 gastroenterologists who were not engaged in the evaluation of the endoscopic scores.

Endoscopic images were evaluated by 6 experts and 20 trainees, who did not consult with each other. Among the raters, the experts were gastroenterologists certified by the Japan Gastroenterological Endoscopy Society who

Table 1. Methods for evaluating endoscopic scores

		Score
MES (Mayo endoscopic subscore)		
Normal or inactive disease		0
Mild disease (erythema, decreased vascular pattern, mild friability)		1
Moderate disease (marked erythema, lack of vascular pattern, friability, erosions)		2
Sever disease (spontaneous bleeding, ulceration)		3
UCEIS (Ulcerative colitis endscopic index of severity)		
Vascular pattern	Normal	0
	Patchy obliteration	1
	Obliterated	2
Bleeding	None	0
	Mucosal	1
	Luminal mild	2
	Luminal moderates or server	3
Erosions and ulcers	None	0
	Erosions	1
	Superficial ulcer	2
	Deep ulcer	3
EAI (Endscopic activity index)		
Size of ulcers	None	0
	erosion/small ulcres	1
	Intermediate	2
	Wide-ranged mucosal defect	3
Deoth of ulcers	None	0
	Shallow	1
	Intermediate	2
	Deep	3
Bleeding	None	0
	Contact bleeding	1
	Spontaneous bleeding	2
	Massive bleeding	3
Mucosal edema	None	0
	Mild	1
	Moderate	2
	Server	3
Redness	None	0
	Mild	1
	Marked	3
Mucous exudate	None	0
	Mild	1
	Marked	2

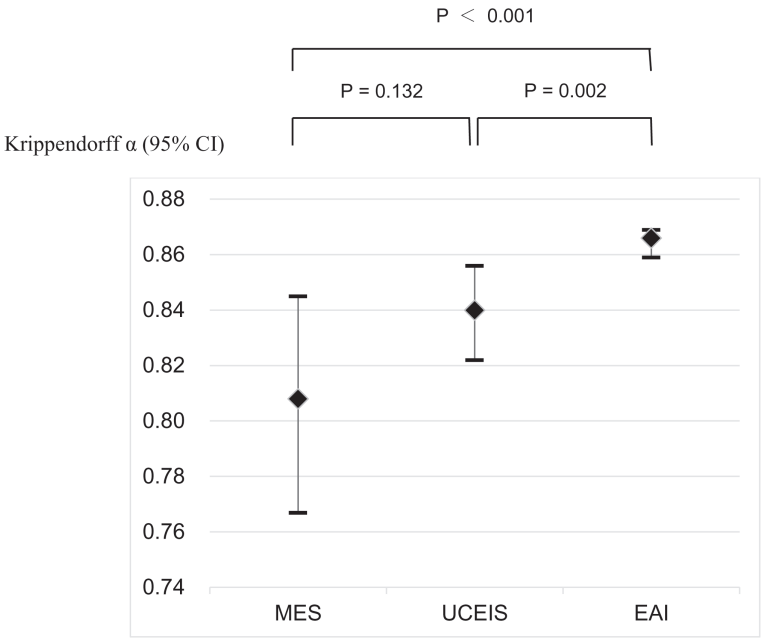
were engaged in the diagnosis and treatment of lower gastrointestinal disease in our hospital and had more than 9 years of experience in colonoscopy. The trainees were gastroenterologists with less than 5 years of experience in colonoscopy who were rotated to the Department of Gastroenterology in our hospital at the time of the evaluation.

Regarding the methods for evaluating endoscopic images in patients with UC, endoscopic scores were evaluated according to the MES, UCEIS, and EAI, using 20 color printed sheets of endoscopic images (Figure 1). The results were described on an evaluation form. The MES scores were classified into 4 categories: 0, normal or inactive disease; 1, mild disease; 2, moderate disease; and 3, severe disease (Table 1).[2] The UCEIS scores were calculated as the sum of the following 3 variables: vascular pattern (0－2 points), bleeding (0－3 points), and erosion and ulcers (0－3 points). The highest UCEIS point score in the present study was 8 (Table 1).[7] The EAI scores were calculated as the sum of 6 items: the size of ulcers (0－3 points), the depth of ulcers (0－3 points), bleeding (0－3 points), mucosal edema (0－3 points), redness (0－2 points), and mucous exudate (0－2 points). The highest EAI point score in the present study was 16 (Table 1).[8] Evaluation was performed by raters who were blinded to the clinical data of patients with UC, such as disease duration, severity, and treatment.

The level of agreement among raters for the MES, UCEIS, and EAI was regarded as the primary variable. The level of agreement for each UCEIS and EAI variable was examined. Because the MES has several indicators for each severity level and is not an independent variable, the level of agreement could not be objectively evaluated. To clarify the influence of endoscopic experience on the level of agreement among the raters, the level of agreement for each score was compared between the experts and the trainees. Regarding the level of agreement within the raters, experts repeatedly evaluated endoscopic images while changing the order more than 6 months

Table 2. Level of agreement among raters for variables in endoscopic scores

	Krippendorff's α
MES	0.808
UCEIS	0.840
Vascular pattern	0.769
Bleeding	0.627
Erosions and ulcers	0.741
EAI	0.866
Size of ulcers	0.751
Depth of ulcers	0.701
Bleeding	0.609
Mucosal edema	0.733
Redness	0.715
Mucous exudate	0.631



95% CI, 95% confidence interval

Figure 2. Comparison of the level of agreement of endoscopic scores among all the raters

after the initial evaluation to compare the level of agreement interobserving raters for the MES, UCEIS, and EAI. Regarding the UCEIS and EAI, the level of agreement for each variable was examined. Concerning the trainees, many gastroenterologists were rotating at associated hospitals at the time of the evaluation so that they could only make one evaluation for each patient. Therefore, the level of agreement of the interobserver raters themselves could not be examined. The ethics committee of our hospital was requested to approve our study. However, as it was a retrospective study of only endoscopic images that did not include any personal information, no ethical review was required.

Statistical analyses

The level of agreement of the intra- and interobserver raters was evaluated with Krippendorff's α coefficients while using responses evaluated according to an ordinal scale. Krippendorff's α coefficients were compared by using a bootstrap method to calculate the 95% confidence interval and P values. P values of <0.05 were considered to indicate statistical significance. Statistical analysis was performed with SSPS, version 23.0 (IBM, Tokyo, Japan).

Results

Level of agreement among the raters

The level of agreement among raters (α coefficient) was 0.808 for the MES, 0.840 for the UCEIS, and 0.866 for the EAI (Table 2). On comparing alpha coefficients for each variable in the UCEIS and EAI, the α coefficient in the UCEIS was 0.627 for bleeding, which was lower than those for vascular pattern and for erosion and ulcers. In the EAI, the α coefficient was 0.609 for bleeding and 0.631 for mucous exudate, which were lower than those for other variables. On comparison of the level of agreement among endoscopic scores, the α coefficient in the EAI was significantly higher than those in the MES and UCEIS (Figure 2). The α coefficient did not differ between the MES and the UCEIS.

In the comparison of the level of agreement (α coefficient) according to endoscopic experience, the α coefficient for each score among the trainees was 0.799 for the MES, 0.824 for the UCEIS, and 0.857 for the EAI (Table 3). The α coefficient was significantly higher in the EAI than that in the MES ($P < 0.001$). The α coefficient for each score among the experts was 0.839 for the MES, 0.891 for the UCEIS, and 0.883 for the EAI, with no significant differences among the scores. In the comparison of the level of agreement for each variable in the UCEIS and EAI, the α coefficients for bleeding and for erosion and ulcers in the UCEIS were significantly higher among the experts. In the EAI, comparison was performed in a similar fashion. No differences were found between the trainees and the experts. On comparing the α coefficients for each score betwen the trainees and the experts, the α coefficients in the UCEIS and EAI were significantly higher among the experts than those among the trainees.

Level of agreement among the interobserver raters

The level of agreement (α coefficient) expert

Table 3. Comparison of the level of agreement of endoscopic scores between trainees and experts

	Krippendorff's α		
	Trainees (n = 20)	Experts (n = 6)	P value
MES	0.799	0.839	0.108
UCEIS	0.824	0.891	<0.001
Vascular pattern	0.761	0.804	0.292
Bleeding	0.593	0.744	<0.001
Erosions and ulcers	0.722	0.822	0.008
EAI	0.857	0.883	0.032
Size of ulcers	0.740	0.788	0.170
Depth of ulcers	0.697	0.715	0.660
Bleeding	0.605	0.632	0.606
Mucosal edema	0.739	0.718	0.526
Redness	0.727	0.696	0.538
Mucous exudate	0.613	0.686	0.296

interobserver raters was 0.886 for the MES, 0.957 for the UCEIS, and 0.954 for the EAI (Table 4). When comparing the α coefficients for each variable in the UCEIS and EAI, the α coefficient for bleeding was lowest in the UCEIS and EAI. On comparison of the level of agreement among endoscopic scores, the α coefficient in the EAI tended to be slightly, but not significantly, higher than that in the MES (Figure 3).

Discussion

Evaluation of colonoscopy is essential for the diagnosis of UC and the assessment of its severity. Evaluation of intestinal lesions with colonoscopy plays an important role in the assessment of the response to drug therapy. In particular, owing to the dissemination of potent drugs such as anti-TNFα antibody preparations and immunosuppressants, not only the disappearance of clinical symptoms but also mucosal healing has been recognized as the treatment goal of UC.[1] Several studies have shown that endoscopic severity at the time of clinical remission is related to outcomes such as the incidence of recurrence.[10-12] In Japan, the number of patients with UC has been increasing, suggesting that opportunities for endoscopists to perform colonoscopy in patients with UC are increasing even among non-specialists of inflammatory bowel disease. Endoscopists who are in charge of colonoscopy for UC are required to accurately and objectively assess the severity of intestinal lesions. In particular, endoscopic scores are often used in the evaluation of colonoscopic findings in clinical trials of new drugs and clinical studies. The use of endoscopic scores allows the severity of intestinal lesions to be assessed as numerical values. In particular, the evaluation of changes in endoscopic scores before and after remission induction therapy may contribute to the objective evaluation of treatment response.

Endoscopic scores used to assess the severity of UC should accurately assess the status of disease, be straightforward, and be highly reproducible. The main aim of the present study was to evaluate and compare the reproducibility of each score in the UCEIS and EAI, in

Table 4. Level of agreement among raters of variables in endoscopic scores

	Krippendorff's α	95% CI	
MES	0.886	0.828	0.999
UCEIS	0.957	0.861	0.974
Vascular pattern	0.934	0.695	0.999
Bleeding	0.818	0.660	0.853
Erosions and ulcers	0.908	0.829	0.971
EAI	0.954	0.908	0.997
Size of ulcers	0.891	0.853	0.999
Depth of ulcers	0.861	0.456	0.995
Bleeding	0.809	0.483	0.932
Mucosal edema	0.901	0.797	0.961
Redness	0.844	0.709	0.941
Mucous exudate	0.853	0.771	0.992

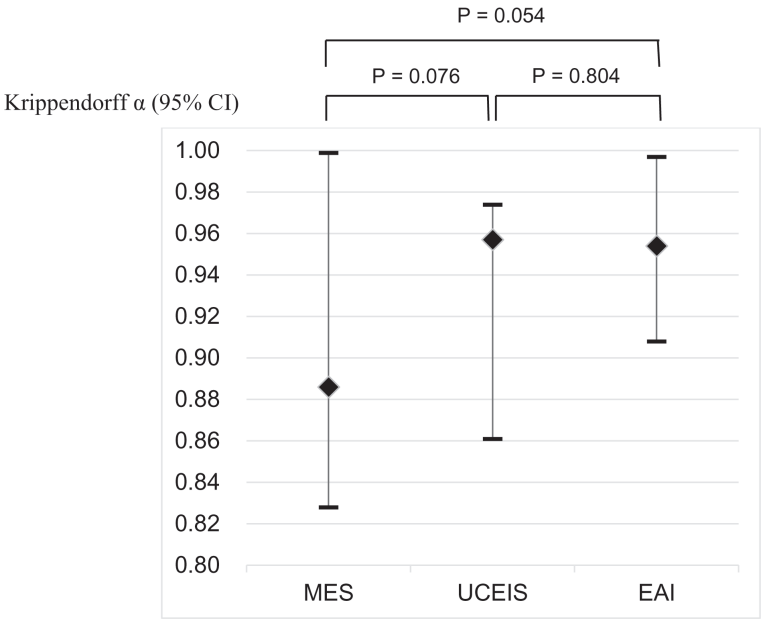


Figure 3. Comparison of the level of agreement of endoscopic scores among the expert raters

addition to the MES, which has been frequently used in clinical trials, among all the endoscopic scores for UC. To assess the reproducibility of endoscopic scores, the level of agreement among several raters and the level of agreement that the same rater evaluated at different time phases (the level of agreement in the interobserver raters) should be examined. In the present study, we evaluated not only the level of agreement among raters but also the level of agreement in the interobserver experts of gastrointestinal endoscopy to compare the 3 types of endoscopic scores.

Regarding the MES and UCEIS among endoscopic scores used in the present study, the level of agreement among the intra- and interobserver raters has been examined previously.[5-7] The level of agreement is not high among raters in the MES.[5,6] Regarding the UCEIS, Travis et al.[7] evaluated the level of agreement among the intra- and interobserver raters and found that the level of agreement among raters was high (κ = 0.96). To our knowledge, regarding the EAI, the reproducibility has not been examined previously. Furthermore, the reproducibility of the 3 types of endoscopic scores used in the present study has not previously been compared. We previously reported that the UCEIS and EAI are more useful than the MES for accurately evaluating the short-term response to drug therapies for UC.[9] To facilitate the use of the UCEIS and EAI in clinical studies of UC, the reproducibility of evaluation using the UCEIS and EAI should be compared with that using the MES, which has been frequently used in clinical trials of new drugs.

Among the 3 types of endoscopic scores evaluated in the present study, the level of agreement among raters was significantly higher in the EAI than in either the UCEIS or the MES, suggesting that the EAI has high reproducibility with the least variation. The EAI has 6 variables and is used with a 3- to 4-step scoring system, which may be the cause of the fact that the level of agreement among raters was high. The level of agreement among raters was lowest in the MES, which might have been caused by the fact that inactive and severe diseases were evaluated using several endoscopic findings as an indicator and that it might have been difficult to differentiate between an erosion and a small ulcer and to assess the friability of the mucosa. In both the EAI and the UCEIS, the level of agreement for bleeding was lowest among the variables. Travis et al.[7] reported that the level of agreement for bleeding was lowest among the UCEIS variables, which was consistent with the results of our study. When evaluating the presence or absence of and the severity of bleeding from the affected area, endoscopy-induced bleeding may occur at the time of endoscope removal. In the UCEIS, the pros or cons of the removal of blood clots by water irrigation are regarded as an indicator of evaluation. Evaluation using a video is required from the reffering physicians. In the present study, however, endoscopic still images were used for evaluation, which may have caused the low level of agreement for the evaluation of bleeding.

To accurately evaluate the reproducibility of endoscopic scores, the influence of the examinars' endoscopic experience should also be evaluated. As endoscopic experience increases, the level of agreement among the raters should accordingly become higher. In the present study, the level of agreement in the UCEIS and EAI was higher among experts than that among trainees. In both of those scores, scoring was performed for each variable. This point was different from that in the MES. As their endoscopic experience increases, the variation of scoring among raters may decrease. The MES, frequently used in clinical studies, is simple, but the scores vary widely among raters. In particular, as the raters' endoscopic experience decreases, the accuracy of the reproducibility becomes lower.[5,6,13] In the present study, the level of agreement among raters in both experts and trainees was lower in the MES than in other endoscopic scores. In trainees, in particular, a significant difference was found in the EAI. On comparison for each variable in the UCEIS and EAI, the level of agreement for the evaluation of bleeding was lowest in both the experts and the trainees. In particular, the level of agreement in the EAI was significantly lower among the trainees than that among the experts. So that trainees can evaluate highly reproducible endoscopic scores, methods may be required whereas mucin and blood clots are removed as much as possible by bowel preparation; and that evaluation should be performed at the time of the colonoscope insertion and, whenever possible, using video. The level of agreement among interobserver raters did not significantly differ among the 3 types of endoscopic scores. However, the level of agreement tended to be higher in the EAI than that in the MES, although no significant difference was observed. This may have been caused due to the evaluation having been done only by experts. Futher studies are warranted to confirm whether or not similar results can be obtained by trainees.

Among endoscopic scores evaluated in the present study, the MES can easily be used to assess the severity of colonoscopic findings. However, we often experience the phenomenon in daily medical practice that the score itself cannot be changed even though endoscopic findings

are being improved. Even if ulcers shrink and became shallower after pharmapseudical therapy, and there is no scarring (Figure 4), the score remains the same even if the MES 3 is not changed. Regarding this point, the UCEIS and EAI adopt a scoring system for each variable and, therefore, are considered to be able to more accurately assess the severity of intestinal lesions in patients with UC. In particular, the EAI has a high level of agreement among raters and is considered the most highly recommended endoscopic score when the severity of intestinal lesions in UC patients is evaluated by several raters with varrying levels of experience.

The limitations of this study were that it was conducted in a single center and that the number of endoscopic images evaluated was only 20. Further studies of many more endoscopic images are warranted for more accurate evaluations. In a study by Travis et al.,[7] endoscopic videos were used to evaluate the UCEIS; however, in the present study, still images were used. This also may have somewhat influenced the results. However, we believe that it is extremely difficult to observe all colonoscopic findings of all patients with UC in videos in regular daily medical pracice because of time restrictions.

These results showed that the EAI is highly reproducible among endoscopic scores used in patients with UC. New drug therapies for UC are continuously approved in clinical practice so that an accurate assessment of treatment response becomes even more important. When placing importance on reproducibility, in the evaluation of the severity of colonoscopic findings, we recommend using the EAI. The EAI is evaluated on the basis of 6 items. However, in daily medical practice, it is important to evaluate colonoscopic findings as easily as is possible with the UCEIS and MES. Thus, the development of a more accurate and universal endoscopic scoring system, and its facilitation, is anticipated and greatly desired.

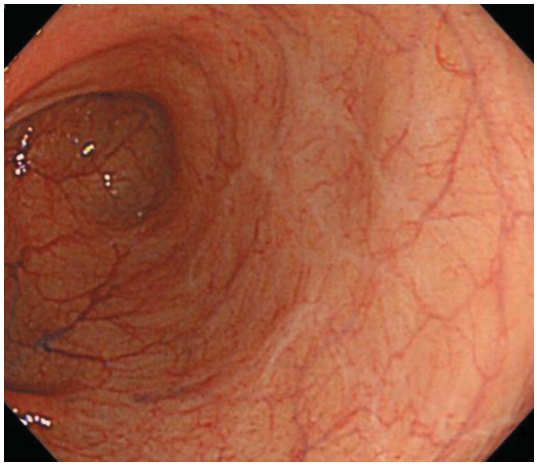


Figure 4. Endoscopic image of a patient who had ulcerative colitis, now healed without scars

Acknowledgements

We thank the participating staffs of the Department of Gastroenterology, and the Research and Development Center for New Medical Frontiers, Kitasato University School of Medicine for their help with image interpretation. We also thank Mr. Peter Star, Medical Network, Tokyo, for advice on the English language of the manuscript.

Conflicts of Interest: None

References

1. D'Haens G, Van Deventer S, Van Hoqezand R, et al. Endoscopic and histological healing with infliximab anti-tumor necrosis factor antibodies in Crohn's disease: A European multicenter trial. *Gastroenterology* 1999; 116: 1029-34.
2. Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. *N Engl J Med* 1987; 317: 1625-9.
3. Rutgeerts P, Sandborn WJ, Feagan BG, et al. Infliximab for induction and maintenance therapy for ulcerative colitis. *N Engl J Med* 2005; 353: 2462-76.
4. Feagan BG, Rutgeerts P, Sands BE, et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med* 2013; 369: 699-710.
5. Osada T, Ohkusa T, Yokoyama T, et al. Comparison of several activity indices for the evaluation of endoscopic activity in UC: inter- and intraobserver consistency. *Inflamm Bowel Dis* 2010; 16: 192-7.
6. Feagan BG, Sandborn WJ, D'Haens G, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology* 2013; 145: 149-57.
7. Travis SPL, Schnell D, Krzeski P, et al. Reliability and initial validation of the ulcerative colitis endoscopic index of severity. *Gastroenterology* 2013; 145: 987-95.
8. Naganuma M, Ichikawa H, Inoue N, et al. Novel endoscopic activity index is useful for choosing treatment in severe active ulcerative colitis patients. *J Gastroenterol* 2010; 45: 936-43.

9. Kawagishi K, Yokoyama K, Kobayashi K. What endoscopic scores are useful for evaluating treatment response in patients with active ulcerative colitis? ECCO Congress, The Netherlands 2014.
10. Carbonnel F, Lavergne A, L?mann M, et al. Colonoscopy of acute colitis. A safe and reliable tool for assessment of severity. *Digest Dis Sci* 194: 39; 1550-7.
11. Daperno M, Sostegni R, Scaglione N, et al. Outcome of a conservative approach in severe ulcerative colitis. *Digest Liver Dis* 2004: 36; 21-8.
12. Yokoyama K, Kobayashi K, Mukae M, et al. Clinical study of the relation between mucosal healing and long-term outcomes in ulcerative colitis. *Gastroenterol Res Pract* 2013; 2013: 192794.
13. Daperno M, Comberlato M, Bossa F, et al. Training programs on endoscopic scoring systems for inflammatory bowel disease lead to a significant increase in interobserver agreement among community gastroenterologists. *J Crohn Colitis* 2017; 11: 556-61.