Reliability and Initial Validation of the Ulcerative Colitis Endoscopic Index of Severity

SIMON P. L. TRAVIS,[1] DAN SCHNELL,[2] PIOTR KRZESKI,[3] MARIA T. ABREU,[4] DOUGLAS G. ALTMAN,[5] JEAN–FRÉDÉRIC COLOMBEL,[6] BRIAN G. FEAGAN,[7] STEPHEN B. HANAUER,[8] GARY R. LICHTENSTEIN,[9] PHILIPPE R. MARTEAU,[10] WALTER REINISCH,[11] BRUCE E. SANDS,[12] BRUCE R. YACYSHYN,[13] PATRICK SCHNELL,[14] CHRISTIAN A. BERNHARDT,[15] JEAN–YVES MARY,[16] and WILLIAM J. SANDBORN[17]

[1]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, England; [2]Statistical Consultant, Middletown, Ohio; [3]Medpace, Warsaw, Poland; [4]Division of Gastroenterology, University of Miami Leonard M. Miller School of Medicine, Miami, Florida; [5]Centre for Statistics in Medicine, University of Oxford, Oxford, England; [6]Hôpital Claude Huriez, Centre Hospitalier Universitaire de Lille, Lille, France; [7]Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, London, Ontario, Canada; [8]Section of Gastroenterology & Nutrition, University of Chicago Medical Center, Chicago, Illinois; [9]Division of Gastroenterology, Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; [10]AP-HP, Hôpital Lariboisière Medicosurgical Department of Digestive Diseases and University Denis Diderot, Paris, France; [11]Universitätsklinik Innere Medizin III, Abteilung Gastroenterologie und Hepatologie, Medical University of Vienna, Vienna, Austria; [12]Dr Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, New York; [13]Division of Digestive Diseases, University of Cincinnati, Cincinnati, Ohio; [14] Mathematics Department Ohio State University, Columbus, Ohio; [15]Bernhardt Regulatory Consulting, Cincinnati, Ohio; [16]INSERM Unité 717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France; and [17]Division of Gastroenterology, University of California San Diego, La Jolla, California

See Covering the Cover synopsis on page 917. See related article, Turner D et al, on page 1140 in *CGH*.

BACKGROUND & AIMS: We studied the reliability of the previously described Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and validated it with an independent cohort of investigators. **METHODS:** We created a new library of 57 videos of flexible sigmoidoscopy and stratified them based on disease severity. Twenty-five investigators were each randomly assigned to assess 28 videos (which included 4 duplicates to assess intraobserver reliability). Investigators were blinded to clinical details except for 2 of 4 duplicated videos (to assess the impact of knowledge of symptoms on assessment). Three descriptors (“vascular pattern”, “bleeding”, and “erosions and ulcers”) comprising the UCEIS were scored with a visual analogue scale (VAS) to assess overall severity. Intrainvestigator and interinvestigator agreement was characterized by κ statistical analysis; reliability ratios were used to compare VAS and UCEIS scores. **RESULTS:** There was a high level of correlation between UCEIS scores and overall assessment of severity (correlation coefficient, 0.93). Internal consistency (Cronbach α analysis) was 0.86. Intrainvestigator and interinvestigator reliability ratios for UCEIS scores were 0.96 and 0.88, respectively. Intrainvestigator agreement in determination of the UCEIS score was good ($\kappa = 0.72$), with individual descriptors ranging from a κ of 0.47 (for bleeding) to 0.87 (for vascular pattern). Interinvestigator agreement in determination of UCEIS scores was moderate ($\kappa = 0.50$), with descriptors ranging from a κ of 0.48 (for bleeding) to 0.54 (for vascular pattern). Intrainvestigator variability in determining UCEIS scores did not change appreciably when a video was presented with clinical details. **CONCLUSIONS: The UCEIS and its components show satisfactory intrainvestigator and interinvestigator reliability. Among investigators, the UCEIS accounted for a median of 86% of the variability in evaluation of overall severity on the VAS when assessing the endoscopic severity of UC and was unaffected by knowledge of clinical details.**

Keywords: Endoscopic Score; Endoscopic Severity; Inflammatory Bowel Disease; Disease Activity.

Endoscopy is extremely valuable for evaluating the efficacy of new treatments for patients with ulcerative colitis (UC).[1,2] However, valid endoscopic scoring systems are needed to standardize end points and facilitate meaningful comparisons.[3] Interobserver variation in endoscopic assessment of disease severity may alter clinical trial outcomes and have a substantial effect on therapeutic or regulatory decisions.[4] Several activity indices for UC incorporate endoscopic data, with the Mayo Clinic Index or the Ulcerative Colitis Disease Activity Index commonly used in trials conducted to seek regulatory approval,[1] but these instruments have not undergone appropriate validation or rigorous reliability assessment.[1,5] To address the need for a highly dependable instrument for assessing the endoscopic severity of UC, we evaluated variations in the overall endoscopic assessment of disease severity, as well as intraindividual and interindividual variations of descriptive terms (“descriptors”), to create the Ulcerative Colitis Endoscopic Index of Severity (UCEIS).[6] The UCEIS was developed in 2 phases: (1) the level of disagreement among investigators and 10 descriptors, each with 3 to 5 levels of severity, was determined and (2) interobserver and intraobserver variability for each descriptor was investigated. A model was then constructed that best represented overall endoscopic severity evaluated on a visual analogue scale (VAS), incorporating 3 descriptors, each with specific

Abbreviations used in this paper: CI, confidence interval; UC, ulcerative colitis; UCEIS, Ulcerative Colitis Endoscopic Index of Severity; VAS, visual analogue scale.

0016-5085
http://dx.doi.org/10.1053/j.gastro.2013.07.024

Table 1. Descriptors, Levels, and Definitions Used as Anchor Points for Evaluating UC

Descriptor (score most severe lesions)	Likert scale anchor points	Definition
Vascular pattern	Normal (0)	Normal vascular pattern with arborization of capillaries clearly defined or with blurring or patchy loss of capillary margins
	Patchy obliteration (1)	Patchy obliteration of vascular pattern
	Obliterated (2)	Complete obliteration of vascular pattern
Bleeding	None (0)	No visible blood
	Mucosal (1)	Some spots or streaks of coagulated blood on the surface of the mucosa ahead of the scope that can be washed away
	Luminal mild (2)	Some free liquid blood in the lumen
	Luminal moderate or severe (3)	Frank blood in the lumen ahead of the endoscope or visible oozing from the mucosa after washing intraluminal blood, or visible oozing from a hemorrhagic mucosa
Erosions and ulcers	None (0)	Normal mucosa, no visible erosions or ulcers
	Erosions (1)	Tiny (≤5 mm) defects in the mucosa of a white or yellow color with a flat edge
	Superficial ulcer (2)	Larger (>5 mm) defects in the mucosa, which are discrete fibrin-covered ulcers when compared with erosions but remain superficial
	Deep ulcer (3)	Deeper excavated defects in the mucosa with a slightly raised edge

NOTE. The worst affected area of the colon visible at sigmoidoscopy was scored. Although the original version of the UCEIS[6] gave a score of 1 to the normal appearance of a descriptor, a collective decision was made to change the numbering of the levels with normality awarded a score of 0, so that the simple sum of the UCEIS ranges from 0 to 8.

definitions: vascular pattern (3 levels), bleeding (4 levels), and erosions and ulcers (4 levels) (Table 1). The worst disease area was scored, and the final score represented the sum of the components, with the UCEIS ranging from 3 (normal) to 11 (most severe). The first 2 phases showed very wide variation in endoscopic interpretation of UC disease severity between specialists but that 3 descriptors with 11 separate levels explained 90% of the variance between observers. After the first 2 phases it was concluded that the UCEIS accurately predicted overall assessment of endoscopic severity of UC, but that it should be assessed for reliability and validated before it could be used as an outcome measure in clinical trials or in routine clinical practice.[6]

In the present study, we performed an independent reliability assessment of the UCEIS with a separate cohort of videos and investigators. The primary objective of the present study was to assess the reliability of UCEIS scoring and perform an initial validation in an independent cohort of videos and investigators after appropriate training. Secondary objectives included an assessment of the impact of endoscopists' knowledge of clinical details on the evaluation of endoscopic disease severity.

Materials and Methods

Terminology

For consistency in the text, the word "index" refers to an instrument for assessing activity, "descriptor" refers to an item within that index with severity allocated on a Likert scale, and "level" refers to the severity graded for an item. "Score" is the overall measure provided by an index.

Development of the UCEIS

Initial development of the UCEIS has been reported.[6] In brief, a library of 670 video sigmoidoscopies from patients with Mayo Clinic scores of 0 to 11, supplemented by 10 videos from 5 people without UC and 5 hospitalized patients with acute, severe UC, was used. Phase 1 mapped inconsistency in overall endoscopic assessment of 16 of 24 video sigmoidoscopies by specialists (the clinical authors) and defined word for word by common agreement 10 endoscopic descriptors that evaluated components of the visual image. Phase 2 was conducted in a separate cohort of 30 investigators from 13 countries. The investigators rated descriptors in 25 of 60 randomly assigned videos and assessed overall endoscopic severity on a VAS from 0 to 100. An index (the UCEIS) consisting of the sum of 3 descriptors, each with 3 or 4 levels of severity, was then constructed that could be tested for reliability and validation (Table 1). Interobserver and intraobserver variations in these descriptors were also quantified. Phase 3 of the study is reported here.

Assessment of the Reliability of the UCEIS

Investigators. Investigators were recruited to reflect a range of geographic and institutional characteristics (see Acknowledgments) from gastroenterologists known to have endoscopic training in trials of inflammatory bowel disease or known to the authors to have an interest in endoscopy and inflammatory bowel disease. Each investigator was then further trained to ensure consistency in understanding and use of the descriptors for assessing endoscopic severity. Training involved assessing video clips of each descriptor at each level, each with an agreed definition of severity. During training, investigators scored 4 standardized videos from phase 2 that included characteristics of the 3 descriptors. To qualify, investigators had to identify correctly the level of the descriptor "erosions and ulcers" on each video and the descriptors "vascular pattern" and "bleeding" within one level of the correct response on each video. Investigators failing to qualify at first assessment were permitted one retest that consisted of correctly scoring 2 of 3 different examples (from different videos) of the descriptor(s) that they had previously incorrectly scored.

Video selection. Videos performed according to a standard procedure[7,8] were selected (by P.K. and B.R.Y.) from a resource of videos from clinical trials of patients with active UC.[8] Subjects had consented to the anonymized presentation of these procedures (EUDRACT 2006-001310-32). Each video comprised a full-length sigmoidoscopy, edited to remove contact friability test images where present, because this technical test had confused earlier assessment. Also included were recordings from subjects (Oxford LREC 536407Q1605/58ORH) without UC

Table 2. Distribution and Allocation of Videos to Investigators in Phase 3

		Mayo Clinic stratum							
	Normal	0	1–2	3–5	6–7	8–9	10–11	Most severe	Total videos
New videos	4	4	8	8	7	8	7	4	50
Videos used in phase 2	1	0	1	1	1	1	1	1	7
Total no. of videos for evaluation	5	4	9	9	8	9	8	5	57
New videos assigned to each investigator	2	3	3	3	3	3	3	2	22
Duplicates of new videos assigned to investigators		Each investigator was assigned 2 videos that duplicated 2 new videos from among these strata							2[a]
Videos used in phase 2 assigned to investigators	1	1	1	1	1	1	1	1	2
Duplicates for determining impact of symptom knowledge	—	—	1	—	1	—	1	—	2

NOTE. "New" refers to videos not previously used in phases 1 and 2.

[a] Each investigator performed 28 evaluations of 22 new videos, 8 of which were 4 videos scored twice. Two evaluations (of a total of 22 videos and 25 investigators = 550) were missing.

during colorectal cancer screening ("normal") and from patients with the most severe UC who had been hospitalized, some before emergency colectomy. All videos were anonymized throughout the study. A library of 57 videos was created and stratified by clinical disease activity using the Mayo Clinic score. Fifty of the videos were new (ie, not previously assessed in phases 1 or 2). Another 7 were repeated as benchmarks, comprising one each from extreme strata (ie, normal or most severe) and 5 with Mayo Clinic scores between 1 and 11.

Video allocation. Each investigator was randomly assigned 28 of 57 videos in randomized order using a set of Latin squares (Table 2). Twenty-six of the 28 videos did not include clinical details. Each investigator was asked to evaluate the most severely affected area. Two duplicates of new videos (Mayo Clinic strata 1–2, 6–7, or 10–11) were provided to evaluate intrainvestigator agreement. Another 2 videos were repeated and supplemented with clinical details (number of stools/day, severity of rectal bleeding, pretreatment or posttreatment status, and physician's global assessment) to evaluate prior knowledge of such clinical details on endoscopic evaluation. Videos were supplied in 3 batches over a 6-week period both to avoid reader fatigue and to optimize memory extinction for duplicated videos. Duplicates were arranged so that the first of any pair was in the first batch and the second was in the third batch.

Video evaluation. Investigators were asked to evaluate the 3 descriptors comprising the UCEIS (Table 1) in the area worst affected at video sigmoidoscopy. In contrast to phase 2,[6] still photographs from the training were provided for reference during evaluation to facilitate reference to the rating standards. A VAS (0–100) rating overall severity was similar to that used for phase 2. The VAS was used as a reference in the absence of a gold standard endoscopic assessment for reasons previously explained.[6] To enable consistent and convenient data entry, investigators were provided with a data capture program designed by one of the authors (P.S.) that could be run simultaneously with video viewing and save responses after each video was scored. Data files were e-mailed to the sponsor after qualification assessments and for each cohort.

Statistical Analysis

Primary objective. The UCEIS was calculated as the simple sum of vascular pattern (scored 0–2), bleeding (scored 0 to 3), and erosions and ulcers (scored 0–3). Thus, the range of possible UCEIS scores was from 0 to 8. (The original version of the UCEIS[6] gave a score of 1 to the normal appearance of a descriptor, which meant that the total score ranged from 3 to 11. A collective decision was made to change the numbering of the levels, such that normality is awarded a score of 0.)

The association between the UCEIS (including the descriptors and the 2 alternative scoring methods) and the evaluation of overall endoscopic severity by the VAS was quantified using Pearson correlation coefficients. Specifically, each investigator's responses for their set of videos were correlated with the mean overall severity (VAS) for those videos, where video means were computed using the responses of all other investigators. These correlations were summarized by median, minimum, and maximum across investigators. Statistical significance was assumed at a level of 0.05 without adjusting for multiple comparisons. Cronbach's coefficient α, using partial correlation coefficients, was calculated for the overall UCEIS score and for the score with one-at-a-time descriptor deletion to evaluate internal consistency in the UCEIS.[9]

Intrainvestigator and interinvestigator agreements for descriptors and the overall UCEIS score were characterized by κ statistics, qualitatively interpreted by Landis and Koch.[10] The standard κ summarizing the exact level of agreement was used for the descriptors. Because the overall UCEIS score represents a 9-level ordinal scale, a weighted κ was used, taking into account close agreement by assigning a weight of 1 for exact agreement, 0.5 for scores that differed by 1 level, and 0 otherwise. Interobserver κ values were calculated by stratifying by investigator pairs and using the common videos they scored but excluding the second scoring of duplicate videos. An average of investigator-pair κ values ("overall κ") was calculated, where the weighting was the inverse of their variance. Intraobserver and interobserver agreement between the overall evaluation of endoscopic severity on the VAS and the UCEIS was assessed by reliability ratios (also known as intraclass correlation coefficients), estimated using mixed-effect linear models. The reliability ratios for inter-investigator agreement were estimated using a model with terms for "investigator," "video," and "error"; additional terms for "investigator-by-video effects" were used to evaluate intra-investigator agreement.[9] Correlation between the UCEIS and overall severity on the VAS, and all interobserver analyses avoided data from the second read of duplicate videos between investigators, and all those where clinical details were provided. Intraobserver analyses, including those for clinical detail/no clinical detail pairs, only used data from duplicate videos.

Secondary objectives. The impact of knowledge of clinical details was evaluated by comparing UCEIS scores and overall severity scores on the VAS within the 50 clinical details/no clinical details pairs. Simple and absolute differences were computed within each pair. *t* tests were used to test for nonzero mean differences. For comparison purposes, these analyses were repeated for the duplicate pairs in which neither video was presented with clinical details. Analysis of variance with terms for investigator and pair type were

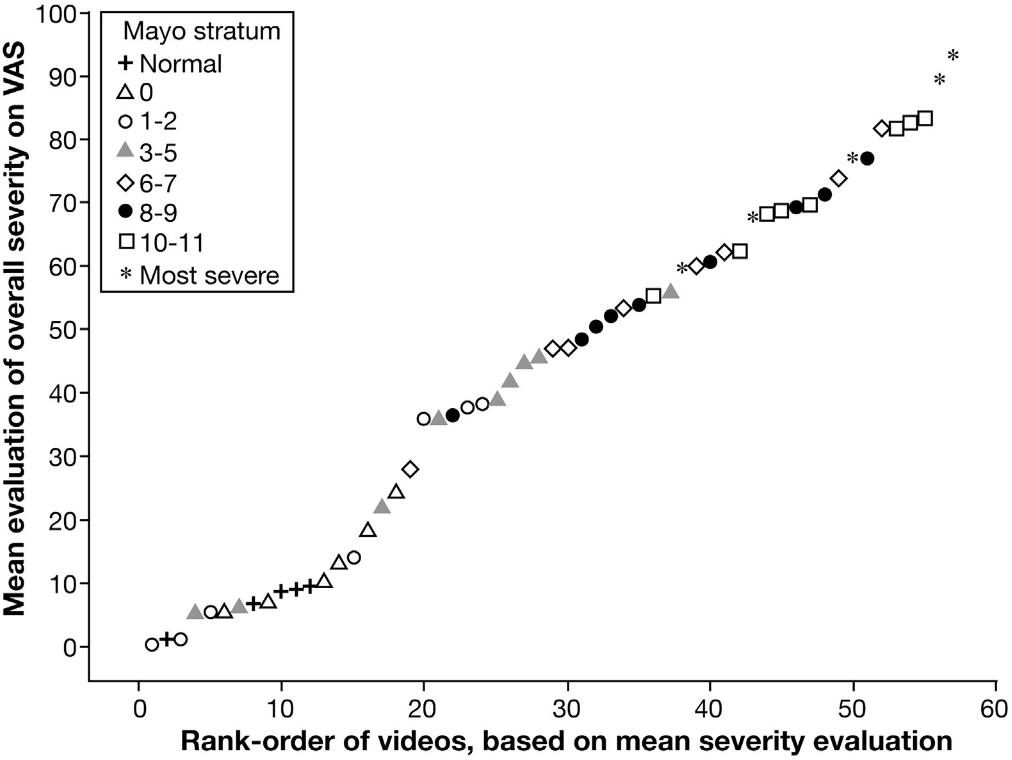

Figure 1. Distribution of overall endoscopic severity assessed by a VAS ranging from 0 to 100.

used to compare absolute differences. Reliability ratios for the UCEIS and overall severity, and intraobserver agreement at the descriptor level, were calculated as described previously. Bowker's test for symmetry[11] tested for presentation order effects (ie, impact of viewing videos with clinical details before or after the blinded version) on responses to descriptors.

Alternative methods for calculation of the UCEIS. Two additional methods for calculating the UCEIS were examined:

1. A normalized sum was used, in which descriptors were combined so as to contribute equally, as one-half "vascular pattern" plus one-third "bleeding" and one-third "erosions and ulcers"; the range of normalized UCEIS scores was then 0 to 3, with 17 possible scores.
2. An alternative version of UCEIS scoring was used to explore an expanded scale should the simple UCEIS prove insensitive to change in the future. Multiple linear regression, with overall severity as the dependent variable, was used to jointly estimate weights for the individual descriptors. Modeling was conducted with investigators treated as fixed effects and descriptors treated as categorical, and as continuous, measures.

Power of differentiation. The design of this study did not permit a direct evaluation of the UCEIS in terms of sensitivity to change between videos at the individual patient level. Nevertheless, the data can be analyzed to assess the power of differentiation across patients (videos). All possible pairings of the 57 videos were formed, for a total of 1596 distinct pairings. Each video was evaluated by between 6 and 15 investigators in the main analysis set. For each pair, mean differences in the UCEIS and overall endoscopic severity on the VAS, and 2-sample *t* tests for differences between videos for evaluation of overall severity on the VAS and the UCEIS were calculated. Proportions of significantly different scores (confirmed by *t* tests) were studied globally and as a function of the difference in endoscopic severity on the VAS.

Comparison with established clinical measures. To compare the UCEIS with established clinical measures for UC, Spearman rank correlation tests were performed between the UCEIS and full Mayo score, partial Mayo score (excluding endoscopic evaluation),[12] stool frequency/rectal bleeding, and patient functional assessment.

Statistical analyses were performed using Statistical Analysis System (SAS, Cary, NC) software version 9.2.

Results

Investigator Qualification

Twenty-nine investigators from 14 countries were screened for participation in the study. Eleven of the 29 succeeded on first qualification and 14 on their second attempt. One investigator failed both times, and 3 were withdrawn due to noncompliance with procedures, resulting in a total of 25 investigators (11 from North America, 9 from Central Europe, and 5 from Western Europe; see Acknowledgments).

Video Evaluation

In total, 698 of the planned 700 evaluations were performed. Each video was assessed by 6 to 15 investigators. The response rate was 100% for assessment of overall severity on the VAS and for all descriptors of these 698 evaluations. The analyses that follow exclude 50 videos from the second evaluation of repeat pairs and the 100 evaluations used for clinical details/no clinical details evaluation, unless stated otherwise.

Range of Disease Severity

Mean overall assessments of endoscopic severity on the VAS ranged from a score of 0.4 for one video in the

Table 3. Description of Evaluations

	Vascular pattern score	Bleeding score	Erosions and ulcers score	UCEIS score[a]	Normalized UCEIS score[b]	VAS score (0–100)
Score						
0	77 (12.9%)	201 (33.6%)	200 (33.4%)	–	–	–
1	183 (30.6%)	211 (35.3%)	139 (23.2%)	–	–	–
2	338 (56.5%)	126 (21.1%)	175 (29.3%)	–	–	–
3	–	60 (10.0%)	84 (14.0%)	–	–	–
Mean (SD)	2.44 (0.71)	2.08 (0.97)	2.24 (1.06)	6.75 (2.45)	1.89 (0.66)	41.26 (30.41)
Median	3.0	2.0	2.0	7.0	2.0	46.0
Minimum, maximum	1.0, 3.0	1.0, 4.0	1.0, 4.0	3.0, 11.0	0.0, 3.0	0, 100

NOTE. n = 598, excluding duplicate videos.

[a] The UCEIS was calculated as a simple sum of the 3 original items: Overall Score = (Vascular Pattern Score) + (Bleeding Score) + (Erosions and Ulcers Score).

[b] The normalized UCEIS score was calculated as a weighted sum of the 3 original items: Normalized UCEIS Score = (Vascular Pattern Score/2) + (Bleeding Score/3) + (Erosions and Ulcers Score/3).

group of Mayo Clinic stratum 1 to 2, to 1.2 to 9.6 for videos in the normal stratum, to 93.4 for a video of the most severe stratum of UC, indicating that the 57 videos embraced the full range of endoscopic UC severity seen in clinical trials and practice (Figure 1). Responses also indicate that the full range of severity was assessed for each descriptor and on the VAS (Table 3).

Initial Validation of the UCEIS

The correlation of the simple sum version of the UCEIS with evaluation of overall severity on the VAS had a median of 0.93 across investigators (minimum, 0.78; maximum, 0.99), indicating that on average the UCEIS captured 86% (derived from 0.93^2) of the variance in investigators' assessments of overall severity. There was also a high level of correlation between the 3 individual descriptors and assessment of overall severity on the VAS: with a median of 0.82 (minimum, 0.55; maximum, 0.90) for vascular pattern, 0.80 (minimum, 0.45; maximum, 0.97) for bleeding, and 0.89 (minimum, 0.78; maximum, 0.96) for erosions and ulcers.

Internal Consistency of the UCEIS

The Cronbach coefficient α was 0.863 for the UCEIS overall (vascular pattern, 0.83; bleeding, 0.80; erosions and ulcers: 0.79). One-at-a-time deletion of descriptors resulted in slightly lower α coefficients (0.79–0.83), indicating that each descriptor contributed positively to the overall UCEIS.

Intrainvestigator and Interinvestigator Agreement for the UCEIS

A total of 50 repeat-pair assessments assessed intraobserver variability. The intrainvestigator reliability ratio for evaluation of overall severity was 0.87 on the VAS and 0.96 for the UCEIS. Intrainvestigator agreement for descriptors ranged from a κ of 0.47 (95% confidence interval [CI], 0.27–0.67) for bleeding to 0.87 (95% CI, 0.74–1.00) for vascular pattern (Table 4), indicating moderate to very good agreement for individual descriptors. The weighted intraobserver κ for the overall UCEIS score was 0.72 (95% CI, 0.61–0.82). A total of 548 video evaluations of 57 videos (22 per investigator, 2 missing; Table 2) assessed interobserver variability. The interinvestigator reliability ratio for overall assessment of severity was 0.78 on the VAS and 0.88 for the UCEIS. Interinvestigator agreement for descriptors ranged from a κ of 0.48 (95% CI, 0.46–0.50) for bleeding to 0.54 (95% CI, 0.50–0.57) for vascular pattern, indicating moderate agreement for individual descriptors between investigators (Table 4). The weighted interobserver κ for the overall UCEIS score was 0.50 (95% CI, 0.49–0.52). In summary, only 4% of the variation in UCEIS scoring in the repeat evaluation data set was attributable to within-investigator variation when scoring the same video twice. Similarly, only 12% of the variation in UCEIS scoring in the main analysis data set was attributable to investigator-to-investigator differences when scoring a common video.

Table 4. Intrainvestigator and Interinvestigator Agreement and Effect of Knowledge of Clinical Details on Intrainvestigator Agreement

Variable	Intrainvestigator agreement (n = 50 pairs of repeat evaluations)	Interinvestigator agreement (n = 548 evaluations)	Clinical details[a]/no clinical details pairs (n = 50 pairs of evaluations)
Vascular pattern	0.87 (0.74–1.00)	0.54 (0.50–0.57)	0.79 (0.63–0.94)
Bleeding	0.47 (0.27–0.67)	0.48 (0.46–0.50)	0.64 (0.47–0.80)
Erosions and ulcers	0.81 (0.67–0.94)	0.53 (0.51–0.57)	0.72 (0.56–0.88)
UCEIS[b]	0.72 (0.61–0.82)	0.50 (0.49–0.52)	0.68 (0.56–0.80)

NOTE. All values are expressed as κ value (95% CI).

[a] Clinical details included age, sex, number of stools/day, severity of rectal bleeding, pretreatment or posttreatment status, and physician's global assessment. Interpretation of standard κ: <0.20, poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, good agreement; 0.81–1.00, very good agreement.[10]

[b] Weighted κ for full 0–8 UCEIS scale; weight = 1.0 for exact agreement, 0.5 for difference of 1 level, and 0 otherwise.

Alternative Methods for UCEIS Scoring

Across investigators, the correlation between the normalized version of the UCEIS and overall severity (VAS) had a median value of 0.94 (minimum, 0.78; maximum, 0.98), very similar to that for the standard UCEIS scoring. The intrainvestigator and interinvestigator reliability ratios were 0.96 and 0.88, respectively, the same as for the 9-level standard UCEIS scoring.

Regression modeling identified an alternative scoring method with unequal descriptor weightings for the UCEIS that also has a high correlation with the overall evaluation of endoscopic severity. Specifically, a weight of 15 applied to the erosions descriptor and 10 to each of the bleeding and vascular pattern descriptors resulted in a UCEIS scale with 18 possible levels and a median (minimum, maximum) correlation across investigators of 0.93 (0.81, 0.99) with overall severity. The intrainvestigator and interinvestigator reliability ratios were 0.96 and 0.88, respectively, the same as for the 9-level standard UCEIS scoring. The Cronbach α for internal consistency decreased from 0.86 to 0.81.

Effect of Clinical Knowledge

The mean difference of UCEIS scores within the 50 clinical details/no clinical details pairs was −0.20 (SD, 0.95; $P = .14$); for overall score (VAS), the mean difference was −1.82 (SD, 15.23; $P = .40$). The absolute differences in UCEIS were 0 or 1 in 45 of the 50 pairs (90%), with a maximum of 4. The mean absolute difference in overall severity was 10.4 (SD, 11.2). The corresponding statistics for the repeat pairs in which neither video had accompanying clinical detail information provided were as follows: mean UCEIS difference of 0.06 (SD, 0.68; $P = .54$), mean overall severity difference of 3.18 (SD, 14.6; $P = .13$), absolute difference in UCEIS of 0 (n = 49; 98%) or 1 (n = 1; 2%), and the absolute difference in mean overall severity of 11.3 (SD, 9.7).

The absolute UCEIS differences within the clinical details/no clinical details pairs did not differ significantly from those within the regular repeated pairs (analysis of variance, $P = .45$) or for overall severity on the VAS (ANOVA, $P = .68$). For the clinical details/no clinical details pairs, the intrainvestigator reliability ratio for evaluation of overall severity was 0.87 on the VAS and 0.93 for the UCEIS. Intrainvestigator agreement with the clinical details/no clinical details pairs was a κ of 0.64 (95% CI, 0.47–0.80) for bleeding, 0.79 (95% CI, 0.63–0.94) for vascular pattern, and 0.72 (95% CI, 0.56–0.88) for erosions and ulcers (Table 4). The weighted κ for the overall UCEIS score within the clinical details/no details pairs was 0.68 (95% CI, 0.56–0.80), very similar to the value of $\kappa = 0.72$ within repeat pairs in which neither video had accompanying clinical details. Viewing the video with clinical details before or after the same video without such did not affect the results ($P > .30$ for all descriptors).

Power of Differentiation

There was a statistically significant difference in mean UCEIS score between videos in 77.3% of the pairings, compared with 71.6% for evaluation of overall severity on the VAS. Figure 2 relates the difference in evaluation of overall severity on VAS between video pairs

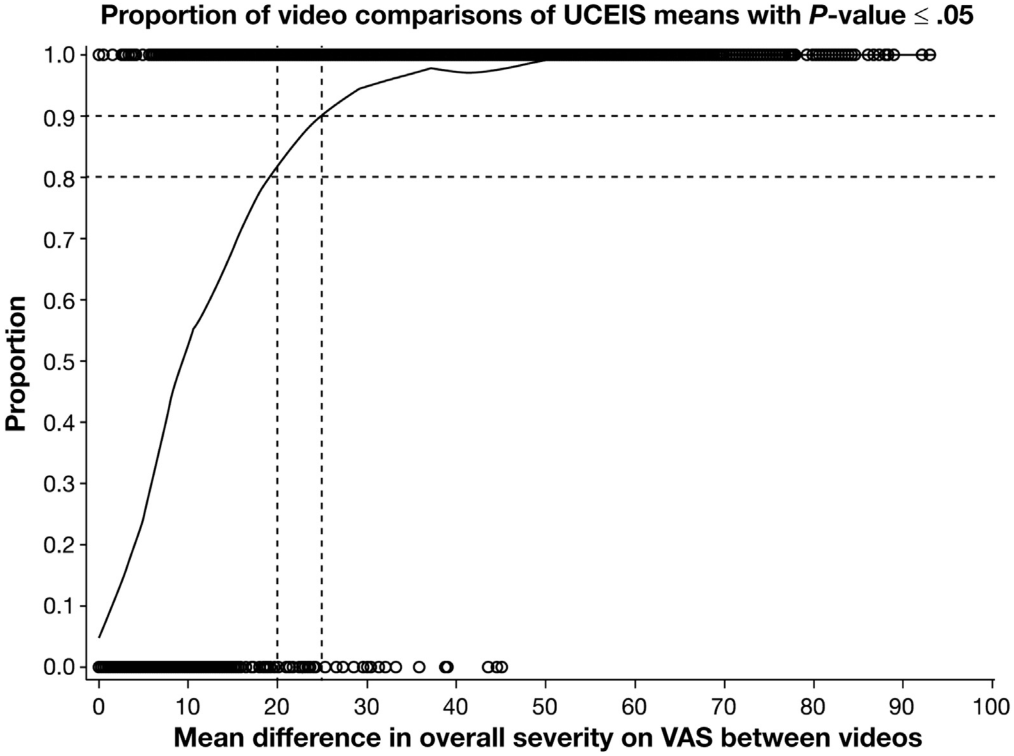

Figure 2. Power of differentiation. The figure relates the mean difference in overall severity between videos measured on the VAS (x-axis) to statistical significance ($P \leq .05$) of the mean difference of UCEIS between videos (y-axis). When the mean difference in overall severity between 2 videos reached 20 units on the VAS, the mean difference in UCEIS between those 2 videos was statistically significant approximately 80% of the time and reached 90% when the mean difference was 25 units on the VAS.

(x-axis) to statistical significance ($P \leq .05$) of the mean difference of the UCEIS between video pairs (y-axis). When the mean difference in overall severity between 2 videos reached 20 units on the VAS, the mean difference in the UCEIS between those 2 videos was statistically significant approximately 80% of the time and reached 90% when the overall difference in severity was 25 units.

Final Version of the UCEIS

The simple sum of different levels of severity was performed as well as a normalized version of calculating the UCEIS, maintaining it as the favored version, with a total score ranging from 0 to 8 (Table 1).

Comparison with Established Clinical Measures

Correlations of the final version of the UCEIS were performed against the full Mayo score, partial Mayo score (excluding endoscopic evaluation), stool frequency/rectal bleeding, and patient functional assessment. Spearman rank correlations ranged from 0.57 (95% CI, 0.51–0.63) for patient functional assessment to 0.73 (95% CI, 0.68–0.77) for the full Mayo score (Table 5).

Discussion

The UCEIS is a reliable instrument for measuring the endoscopic disease activity of UC. After initial assessment for validity, it also appears to be valid, but additional validity testing is needed. Just 3 descriptors (each with 3 or 4 levels of severity) accounted for 86% of the variance in the overall assessment of endoscopic severity. Given the enormous variance in assessment between specialists in the initial evaluation,[6] this represents substantial progress. Correlation of the UCEIS with established UC activity scores was shown to be moderate (stool frequency/rectal bleeding: 0.67 [95% CI, 0.61–0.72]; patient functional assessment, 0.57 [95% CI, 0.51–0.63]) or strong (Mayo score, 0.73 [95% CI, 0.68–0.77]; partial Mayo score, 0.70 [95% CI, 0.64–0.74]). This provides additional support for the performance of the UCEIS using just 3 descriptors (Table 5).

Mean overall assessments of endoscopic severity indicated that the 57 videos, evaluated by an independent cohort of 25 investigators from 14 countries (more than half of whom came from North America or Western Europe), were representative of the full range of endoscopic UC severity seen in clinical practice. Internal consistency (Cronbach coefficient α of 0.86) was good-excellent (ie, >0.70) for the descriptors in the index.[11] Across investigators, correlation between the overall evaluation of endoscopic severity on the VAS and the UCEIS was exceptionally high (median Pearson correlation coefficient of 0.93). The lack of a true gold standard for assessing endoscopic severity of UC was an inevitable shortcoming of the study, so the overall severity assessed on the VAS was used as a reference. It is conceivable that correlation was enhanced by contemporary scoring of both descriptors and the VAS, but the lack of a training calibration for scoring the VAS would have detracted from the correlation. Nevertheless, potential bias was examined by correlating each investigator's responses for their set of videos with the mean VAS for those videos, computed from the responses of all other investigators. Median correlations ranged from 0.80 to 0.93, which suggests that the UCEIS is likely to be a valid assessment of endoscopic severity.

Table 5. Spearman Rank Correlations With the Final Version of the UCEIS

Comparator	Correlation	95% CI
Full Mayo score	0.73	(0.68–0.77)
Partial Mayo score	0.70	(0.64–0.74)
Stool frequency and rectal blood	0.67	(0.61–0.72)
Patient functional assessment	0.57	(0.51–0.63)

NOTE. These analyses exclude the "normal" and "most severe" subjects for which the Mayo and patient functional assessment data were not available.

Intrainvestigator and interinvestigator reliability ratios for the UCEIS were 0.96 and 0.88, respectively, each better than overall severity as measured by the VAS. Intra-observer agreement for each descriptor was moderate to very good (κ of 0.47 [95% CI, 0.27–0.67] for bleeding to 0.87 [95% CI, 0.74–1.00] for vascular pattern) and good for the overall UCEIS score (weighted κ of 0.72 [95% CI, 0.61–0.82]). Interinvestigator agreement was rated as moderate for all descriptors and moderate for the 9-level UCEIS as a whole (weighted κ of 0.50 [95% CI, 0.49–0.52]). It may seem surprising that scoring of bleeding was most subject to variation by the same observer. This may have been the result of investigators' misinterpretation of the descriptions used to define the level of bleeding. Alternatively, this variation may be because investigators did not appreciate the importance of scoring bleeding during insertion of the flexible sigmoidoscope, despite being directed to do so to avoid confusion with contact bleeding. Importantly, however, there was no significant difference in κ statistics between descriptors. Indeed, it is remarkable that this was the only unexpected result in a study notable for a good level of consistency.

Our data suggest that the key to consistent evaluation of endoscopic severity between observers is a standardized system of description. Training is another component. Other work has reported that scores for interobserver and intraobserver weighted κ statistics using established indices are all lower for trainee endoscopists than for specialists, indicating that assessment of disease activity benefits from experience.[13] Assessment of a total of 28 videos could therefore be subject to a training effect, which might bias findings in later assessments. To limit such bias, all investigators underwent initial training and qualification, the order of all videos (including duplicates) was randomized, and the videos were provided in 3 separate batches separated by time to optimize memory extinction between video reading sessions.

Nevertheless, there were anomalies. Normal videos received a higher mean VAS score than those from some patients (Figure 1), although a normal endoscopy is entirely consistent with UC in remission and this must reflect variation around normality. The more important point is that 25 independent investigators evaluated 57 endoscopies and that the range of overall severity on a scale from 0 to 100 was 0.4 to 93.4, indicating that the selected endoscopies gave as wide a range of severity for assessment as reasonably possible. It is conceivable that physician knowledge of clinical information might influence endoscopic assessment.[14] For the UCEIS, knowledge of symptoms had a modest effect overall, although, as might be expected, this had the greater effect on the bleeding descriptor. In the 50 repeat pairs of videos, agreement in the rectal bleeding score between the 2 readings improved from a κ of 0.47 (95% CI, 0.27–0.67) to 0.64 (95% CI, 0.47–0.80) (Table 4) when symptoms were known, but the numbers are small. It is understandable that if symptoms of rectal bleeding are present, then the threshold for describing bleeding at endoscopy (and therefore variability in that description) is reduced. Further evaluation of the impact of clinical details on endoscopists' assessment in UC is warranted.

Sensitivity to change is a valuable property of an index and is best achieved by comparing the delta change to the assessment of variance in patients unchanged after treatment of known efficacy. This needs to be assessed in a prospective clinical trial, although statistical analysis in the current cohort showed that the UCEIS was significantly different 90% of the time when the overall severity on the VAS differed by 25 points. The clinical relevance of this modeling must be regarded as uncertain. In a further step toward its place in research, training, and clinical practice, the UCEIS is currently undergoing development by the European Crohn's and Colitis Organisation as a training tool.

This study is a first step in the validation of the UCEIS. It has confirmed the reliability of the UCEIS, even if further validation is needed to establish thresholds for remission, the clinical relevance of different UCEIS scores, and responsiveness of the UCEIS to change in disease status. The UCEIS is based on evaluation of the most severely affected area at flexible sigmoidoscopy. It is as yet unclear how an overall score might be affected by full colonoscopy or whether it might be applied in colonic segments.[15,16] Colonoscopy could result in a higher UCEIS than sigmoidoscopy simply because a larger area is examined; because scoring is applied to the area of maximum severity, if that area lies proximal to the rectosigmoid colon, the score would increase de facto. This might, in turn, alter the overall evaluation of endoscopic severity. The UCEIS showed consistency in endoscopic evaluation and, if it can be shown to correspond with histological disease activity or validated biomarkers, may facilitate the use of smaller sample sizes in clinical trials due to increased statistical power derived from this consistency. If the UCEIS can demonstrably affect decision making or predict clinical outcome, then this will amplify its role in clinical practice.

The UCEIS reliably evaluates the overall endoscopic severity of UC and accounts for 88% of the variance between endoscopists. It is simple to use, based on the sum of 3 descriptors with a score ranging from 0 to 8. The thresholds for severity and remission remain to be defined, as does the responsiveness to change. In conjunction with a training package to protect the reliability of scoring, it is ready to be further evaluated in clinical trials.

References

1. D'Haens G, Sandborn WJ, Feagan BG, et al. A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. Gastroenterology 2007;132:763–786.
2. Danese S, Fiocchi C. Ulcerative colitis. N Engl J Med 2011; 365:1713–1725.
3. Cooney RM, Warren BF, Altman DG, et al. Outcome measurement in clinical trials for ulcerative colitis: toward standardisation. Trials 2007;8:17–25.
4. Travis SPL, Cooney R, Dunmon P, et al. Conduct of clinical trials in ulcerative colitis: impact of independent scoring of endoscopic severity on results of a randomised controlled trial with a peptide and 5ASA. Am J Gastroenterol 2006;101(Suppl 9):S429.
5. 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;45:936–943.
6. Travis SPL, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Gut 2012;61:535–542.
7. Travis SPL, Yap LM, Hawkey CJ, et al. RDP58 is a novel and potentially effective oral therapy for ulcerative colitis (UC): results of parallel prospective, multicenter, blinded, placebo-controlled trials. Inflamm Bowel Dis 2005;11:713–719.
8. Sandborn WJ, Regula J, Feagan BG, et al. Delayed-release oral mesalamine 4.8 g/day (800-mg tablet) is effective for patients with moderately active ulcerative colitis. Gastroenterology 2009;137: 1934–1943.e1–3.
9. Streiner DL, Norman GR. Health measurement scales. 2nd ed. Oxford University Press, 1995.
10. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363–374.
11. Bowker AH. A test for symmetry in contingency tables. J Am Stat Assoc 1948;43:572–574.
12. Lewis JD, Chuai S, Nessel L, et al. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis 2008;14:1660–1666.
13. 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;16:192–197.
14. Bushnell CD, Goldstein LB. Physician knowledge and practices in the evaluation of coagulopathies in stroke patients. Stroke 2002; 33:948–953.
15. Thia KT, Loftus EV Jr, Pardi DS, et al. Measurement of disease activity in ulcerative colitis: interobserver agreement and predictors of severity. Inflamm Bowel Dis 2011;17:1257–1264.
16. Samuel S, Bruining DH, Loftus EV, et al. Validation of the Ulcerative Colitis Colonoscopic Index of Severity and its correlation with disease activity measures. Clin Gastroenterol Hepatol 2013;11:49–54.

Received March 8, 2013. Accepted July 13, 2013.

Reprint requests

Address requests for reprints to: Simon P. L. Travis, DPhil, FRCP, MBBS, Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford OX3 9DU, England. e-mail: simon.travis@ndm.ox.ac.uk; fax: (44) 1865 858763.

Acknowledgments

The authors thank the investigators who evaluated video endoscopies from North America (David Armstrong, Hamilton, Ontario, Canada; Frank Anderson, Vancouver, British Columbia, Canada; Ronnie

Cyzner, Charlotte, NC; Barry Schneider, Davidson, NC; Tedd Cain, Milwaukee, WI; Alvin Cohen, Hollywood, FL; Brian Sullivan, Chesapeake, VA; Lev Ginzburg, Great Neck, NY; Michael Cohen, Arlington Heights, IL; Christine Lewis, Hagerstown, MD; Tawfik Chami, Zephyrhills, FL), Central Europe (Elena Mikhailova, Gomel, Belarus; Marko Banic, Zagreb, Croatia; Tibor Szaloki, Vac, Hungary; Aldis Pukitis, Riga, Latvia; Limas Kupcinskas, Kaunas, Lithuania; Jacek Sobocki, Cracow, Poland; Michal Kaminski, Warsaw, Poland; Irina Kholina, St Petersburg, Russia; Ivan Jovanovic, Belgrade, Serbia), and Western Europe (Jean-Baptiste Chevaux, Nancy, France; Markus Frenz, Bremen, Germany; Gionata Fiorino, Rozzano, Italy; Ana Ignjatovic, London, England; Daniel Burger, Oxford, England) as well as Anita Hinds and Nicola Deane (Warner Chilcott, Clinical data management), and Francesc Miras Rigol (RPS, programming) for help and support. Biostatistical advice was independent and provided by the sponsors of the study (Warner Chilcott). UCEIS copyright is registered to Warner Chilcott on the principle that there is unrestricted access to the UCEIS.

Conflicts of interest

The authors disclose the following: All authors gave freely of their time and have received no remuneration for the development, reliability assessment, and initial validation of this index. Simon P. L. Travis has in the past received consulting fees from AbbVie, Cosmo Technologies, Ferring Pharmaceuticals, GlaxoSmithKline, Janssen Pharmaceutical Research & Development, Merck, Novartis, Novo Nordisk, Pfizer, Santarus, Schering-Plough, Shire Pharmaceuticals, Sigmoid Pharma, Tillotts Pharma AG, UCB Pharma, Vifor, and Warner Chilcott UK; research grants from AbbVie, Janssen Pharmaceutical Research & Development, Novartis, Pfizer, and UCB Pharma; and payments for lectures/speaker's bureau from AbbVie, Ferring Pharmaceuticals, Merck, Sanofi, and Tillotts Pharma AG. Piotr Krzeski was an employee of Warner Chilcott at the time of this study. Maria T. Abreu has served as a consultant for Abbott Laboratories, Bristol-Myers Squibb, Hospira, Merck, Pfizer, Sanofi-Aventis, and UCB. Jean-Frédéric Colombel has served as consultant, advisory board member, or speaker for AbbVie, Amgen, Bristol-Meyers Squibb, Celltrion, Ferring Pharmaceuticals, Genentech, Giuliani SpA, Given Imaging, Merck, Millennium Pharmaceuticals, Nutrition Science Partners, Pfizer, Prometheus Laboratories, Sanofi, Schering-Plough, Takeda, Teva Pharmaceuticals, UCB Pharma, Vertex, and Dr August Wolff GmbH & Co. Brian G. Feagan has served as consultant, advisory board member, or speaker or provided grant/research support for Abbott Laboratories (AbbVie), ActoGeniX, Albireo Pharma, Amgen, AstraZeneca, Athersys, Avaxia Biologics, Axcan, Boehringer-Ingelheim France, Bristol-Myers Squibb, Celgene, Centocor, CombinatoRx, Elan/Biogen, Ferring Pharma A/S, Genentech, glcare Pharma, Gilead, Given Imaging, GlaxoSmithKline, Ironwood Pharma, Johnson & Johnson/Janssen, Merck, Millennium Pharmaceuticals, Nektar, Novo Nordisk, Novartis, Prometheus Therapeutics and Diagnostics, Prometheus Laboratories, Pfizer, Protein Design Labs, Salix Pharmaceuticals, Serono, Shire Pharmaceuticals, Sigmoid Pharma, Synergy Pharmaceuticals, Takeda, Teva Pharmaceuticals, Tillotts, UCB Pharma, Unity Pharmaceuticals, Warner Chilcott, Wyeth, Zealand Pharma, and Zyngenia. Gary R. Lichtenstein has served as a consultant, provided research support, or received honoraria from Abbott Laboratories (AbbVie), Alaven Pharmaceuticals, Bristol-Myers Squibb, Janssen Ortho Biotech, Elan, Ferring Pharmaceuticals, Hospira, Ironwood, Millennium Pharmaceuticals, Ono Pharmaceuticals, Pfizer, Prometheus Laboratories, Salix Pharmaceuticals, Santarus, Schering-Plough, Shire Pharmaceuticals, Takeda, UCB Pharma, and Warner Chilcott. Philippe R. Marteau has received honoraria for consultancy of or lectures and travel fees from AbbVie, Biocodex, Danone, Dr Falk, Ferring, Ipsen, Merck Médication Familiale, MSD, Nestlé, and Vifor Pharma. Walter Reinisch has served as speaker, consultant, and/or advisory board member for Abbott Laboratories, AbbVie, Aesca, Amgen, Astellas, AstraZeneca, Biogen IDEC, Bristol-Myers Squibb, Cellerix, ChemoCentryx, Celgene, Centocor, Danone Austria, Elan, Ferring Pharmaceuticals, Galapagos, Genentech, Grünenthal, Johnson & Johnson, Kyowa Hakko Kirin Pharma, Lipid Therapeutics, Millennium Pharmaceuticals, Mitsubishi Tanabe Pharma, MSD, Novartis, Ocera, Otsuka, PDL BioPharma, Pharmacosmos, Pfizer, Procter & Gamble, Prometheus Laboratories, Robarts Clinical Trials, Schering-Plough, SetPoint Medical, Shire Pharmaceuticals, Takeda, Therakos, TiGenix, UCB, Vifor, Yakult, Zyngenia Austria, and 4SC. Christian A. Bernhardt received payment from Warner Chilcott for the time needed to support the study logistics (design, protocol development, execution, data analysis and conclusions, and manuscript development). He was employed by Procter & Gamble when the studies of Asacol and RDP-58 were conducted and the endoscopy videos were collected. William J. Sandborn has received consulting fees from Abbott Laboratories (AbbVie), ActoGeniX NV, AGI Therapeutics, Alba Therapeutics, Albireo, Alfa Wasserman, Amgen, AM-Pharma BV, Anaphore, Astellas, Athersys, Atlantic Healthcare, Aptalis, BioBalance, Boehringer-Ingelheim, Bristol-Myers Squibb, Celgene, Celek Pharmaceuticals, Cellerix SL, Cerimon Pharmaceuticals, ChemoCentryx, CoMentis, Cosmo Technologies, Coronado Biosciences, Cytokine PharmaSciences, Eagle Pharmaceuticals, EnGene, Eli Lilly, EnteroMedics, Exagen Diagnostics, Ferring Pharmaceuticals, Flexion Therapeutics, Funxional Therapeutics, Genzyme, Gilead Sciences, Given Imaging, GlaxoSmithKline, Human Genome Sciences, Ironwood Pharmaceuticals, Janssen Pharmaceutical Research & Development, KaloBios Pharmaceuticals, Lexicon Pharmaceuticals, Lycera, Meda Pharmaceuticals, Merck Research Laboratories, Merck Serono, Millennium Pharmaceuticals, Nisshin Kyorin Pharmaceuticals, Novo Nordisk, NPS Pharmaceuticals, Optimer Pharmaceuticals, Orexigen Therapeutics, PDL BioPharma, Pfizer, Procter and Gamble, Prometheus Laboratories, ProtAb, PurGenesis Technologies, Relypsa, Roche, Salient Pharmaceuticals, Salix Pharmaceuticals, Santarus, Schering-Plough, Shire Pharmaceuticals, Sigmoid Pharma, Sirtris Pharmaceuticals, SLA Pharma UK, Targacept, Teva Pharmaceuticals, Therakos, Tillotts Pharma AG, TxCell SA, UCB Pharma, Viamet Pharmaceuticals, Vascular Biogenics, Warner Chilcott UK, and Wyeth; has received research grants from Abbott Laboratories, Bristol-Myers Squibb, Genentech, GlaxoSmithKline, Janssen Pharmaceutical Research & Development, Millennium Pharmaceuticals, Novartis, Pfizer, Procter and Gamble, Shire Pharmaceuticals, and UCB Pharma; has received payments for lectures/speaker's bureau from Abbott Laboratories, Bristol-Myers Squibb, and Janssen Pharmaceutical Research & Development; and holds stock/stock options in EnteroMedics. The remaining authors disclose no conflicts.

Funding

This study was funded by Warner Chilcott. The sponsors paid for video image preparation, logistical support, statistical programming, and the time spent by independent gastroenterologists to evaluate the UCEIS. Editorial support was provided by Strategen Ltd and funded by Warner Chilcott.