ORIGINAL ARTICLE

Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS)

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] Marc Lémann,[9] Gary R Lichtenstein,[10] Phillippe R Marteau,[11] Walter Reinisch,[12] Bruce E Sands,[13] Bruce R Yacyshyn,[14] Christian A Bernhardt,[15] Jean-Yves Mary,[16] William J Sandborn[17]

► Additional files are available online only. To view these files please visit the journal online (http://gut.bmj.com/content/61/4.toc).

For numbered affiliations see end of article.

Correspondence to
Dr Simon P L Travis, Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford OX3 9DU, UK; simon.travis@ndm.ox.ac.uk

Revised 28 August 2011
Accepted 30 August 2011
Published Online First 13 October 2011

ABSTRACT

Background Variability in endoscopic assessment necessitates rigorous investigation of descriptors for scoring severity of ulcerative colitis (UC).

Objective To evaluate variation in the overall endoscopic assessment of severity, the intra- and interindividual variation of descriptive terms and to create an Ulcerative Colitis Endoscopic Index of Severity which could be validated.

Design A two-phase study used a library of 670 video sigmoidoscopies from patients with Mayo Clinic scores 0—11, supplemented by 10 videos from five people without UC and five hospitalised patients with acute severe UC. In phase 1, each of 10 investigators viewed 16/24 videos to assess agreement on the Baron score with a central reader and agreed definitions of 10 endoscopic descriptors. In phase 2, each of 30 different investigators rated 25/60 different videos for the descriptors and assessed overall severity on a 0—100 visual analogue scale. κ Statistics tested inter- and intraobserver variability for each descriptor. A general linear mixed regression model based on logit link and β distribution of variance was used to predict overall endoscopic severity from descriptors.

Results There was 76% agreement for 'severe', but 27% agreement for 'normal' appearances between phase I investigators and the central reader. In phase 2, weighted κ values ranged from 0.34 to 0.65 and 0.30 to 0.45 within and between observers for the 10 descriptors. The final model incorporated vascular pattern, (normal/patchy/complete obliteration) bleeding (none/mucosal/luminal mild/luminal moderate or severe), erosions and ulcers (none/erosions/superficial/deep), each with precise definitions, which explained 90% of the variance (pR^2, Akaike Information Criterion) in the overall assessment of endoscopic severity, predictions varying from 4 to 93 on a 100-point scale (from normal to worst endoscopic severity).

Conclusion The Ulcerative Colitis Endoscopic Index of Severity accurately predicts overall assessment of endoscopic severity of UC. Validity and responsiveness need further testing before it can be applied as an outcome measure in clinical trials or clinical practice.

Significance of this study

What is already known about this subject?
- There is wide variation in the endoscopic assessment of the severity of ulcerative colitis.
- There is no validated instrument.

What are the new findings?
- An index—the Ulcerative Colitis Endoscopic Index of Severity—with three descriptors (vascular pattern, bleeding and ulceration) has been developed that captures 90% of the variance of the overall assessment of endoscopic severity.
- The three descriptors are each graded in three or four levels with precise definitions.
- Friability is excluded from the endoscopic description of severity.

How might this impact on clinical practice?
- Once independently validated, the Ulcerative Colitis Endoscopic Index of Severity will be available for clinical trials, training and practice.

INTRODUCTION

Endoscopy plays a pivotal role in the evaluation of ulcerative colitis (UC). At least nine different scoring systems are used as outcome measures in clinical trials, and endoscopy has an important role in most.[1 2] Indices are typically composite measures that include assessment of symptom severity, quality of life, laboratory tests and endoscopic findings. However, the contribution of endoscopy is index-specific. In the widely used Mayo Clinic index,[2] endoscopy is one of four criteria and just one of two criteria (in addition to rectal bleeding) currently used by the Food and Drug Administration for defining remission. Consequently, interobserver variation in assessing endoscopic activity is important, because disagreement can alter the proportion of patients defined as in remission and influence regulatory decisions.

The original endoscopic grading of UC (Baron index, 1964) was developed before index methodology was defined.[3] It has been used, nevertheless,

in most trials of active UC to this day, with only minor and unvalidated modification.[2] Data supporting the index are scant. It was created by scoring seven endoscopic descriptors in 60 patients, by three observers using rigid sigmoidoscopes. The κ statistic, a measure of interobserver agreement, was not calculated and there was 40% disagreement when grading normal, mild, moderate or severe activity. Efforts have since been made to standardise endoscopic assessment[3] by using the presence of mucosal friability to discriminate between mild (Baron level 1) and moderately active (Baron level 2) disease.[4–6]

Variation between observers in categorising endoscopic disease activity is widely suspected,[1 2 7–10] so the need for this to be quantified appears self-evident. The aims of this study were first to substantiate variation in endoscopic assessment of activity in UC, then to evaluate intra- and interindividual variation of descriptive terms and, finally, to create an Ulcerative Colitis Endoscopic Index of Severity (UCEIS) which could be validated.

METHODS

The study included two phases. Phase 1 mapped inconsistencies in endoscopic assessment and defined the most dependable descriptive terms ('descriptors'). Phase 2 quantified inter- and intraobserver variation in these descriptors, in order to construct an index (UCEIS) that could be validated. 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' is used to refer to the severity graded for an item. 'Score' is the overall measure provided by an index. Common usage has often confused these terms, but they are used as consistently as possible in this paper.

Phase 1

Ten specialists in inflammatory bowel disease (IBD, the authors) graded videos of flexible sigmoidoscopy according to their own practice, in the absence of clinical information. Twenty-four representative videos were selected to represent the widest range of UC activity, guided by the Mayo Clinic score (by PK and BRY) from a library of 670 videos recorded in a standard manner during clinical trials for the treatment of moderately active UC[6] (EUDRACT 2006-001310-32). Within each Mayo Clinic score stratum, consecutive videos were reviewed by one of the co-authors for image quality. Satisfactory quality recordings (sharp image, sufficient bowel preparation) were selected. Videos from fibreoptic endoscopes were discarded. Sixteen videos represented the complete range of severity; 24 videos enabled choice from additional videos in the mid-range of severity, most likely to be affected by interobserver variation. Each investigator was randomly assigned 16 of the 24 videos in randomised order using a set of Latin squares: a core set of eight videos that all investigators evaluated (two for each Baron score) and eight of 16 non-core videos, This kept the number of evaluations by each investigator in the 2–3 h session to a manageable number (16), while still having a common core set (8) and a broad overall pool of videos (24). Investigators were explicitly advised not to apply the Baron index themselves, to avoid biasing their overall assessment of severity in relation to this index. To assess potential scoring differences based on the length of the video,[11] each investigator had two pairs that were shortened from 10–15 min to approximately 5 min, giving a total of 18 videos for each investigator to view. Descriptors of endoscopic severity were selected from previous studies.[3 8 9 12 13] Investigators recorded the presence or absence of 11 descriptors. Overall severity was assessed on a visual analogue scale (VAS, between 0=completely normal and 100=worst ever seen).

To substantiate variability in endoscopic assessment, the level of the Baron index derived from the assessments by investigators was compared with the level assigned by the central reader in the original trial.[7] The precise wording of definitions and video clips illustrating anchor points on three-, four- or five-point Likert scales of severity for each descriptor, were subsequently agreed by consensus during a video teleconference between investigators (table 1).

Phase 2

Fifty core videos were assembled, 40 from the library of 670 videos (by PK and BRY, representing Mayo Clinic levels (scores) 0–11, different from those selected for phase 1), representing six expected severity strata (note selection criteria for phase 1). A further five from individuals without UC and five from patients hospitalised with severe UC who had consented to their anonymised images being used for study (Oxford LREC 536407Q1605/58ORH), represented two additional strata at the expected extremes of endoscopic severity. The five patients admitted with biologically severe UC represented the most severe end of the spectrum of UC, although only 2/5 came for colectomy (one within 6 h of flexible sigmoidoscopy). To evaluate the Contact Friability Test (CFT), 10 different videos representing Mayo Clinic levels 1–11, two for each stratum, were amended to exclude CFT sequences and paired with the complete video showing the CFT.

Each of 30 new investigators from 13 countries, including 19 from the USA and Canada (see 'Acknowledgements') scored 25 videos selected from the 60 recordings, but blinded to clinical information or Mayo Clinic score. Endoscopists were recruited to reflect a range of geographical and institutional characteristics, from investigators with endoscopic training in trials of IBD or known to the authors as having an interest in endoscopy or IBD (840 median colonoscopies and flexible sigmoidoscopies/year (range 100–2100), median 25 years' endoscopy experience, range 8–35). Each investigator was randomly allocated a CDROM containing 15/40 core videos from the library (comprising two to three videos selected from each of the six Mayo Clinic levels), two out of five normal videos from people without UC and two out of five videos from patients with severe UC, together with two out of 10 CFT+/− pairs (table 2). To evaluate intraobserver consistency, each investigator also scored two of their 13 core videos representing Mayo Clinic levels 1–11 twice, in random order. Investigators were asked to score each video using every descriptor in table 1 and to assign an overall assessment of severity using an electronic 0–100 VAS.

Videos were assigned to investigators using an incomplete block design, stratified by expected severity stratum. This randomisation process meant that each video was scored by 10–12 investigators, except for the four videos in Mayo Clinic level 0 stratum, which were each viewed by 15 investigators. Owing to an assignment error, 5/30 investigators were assigned only one and not two normal videos. The order of endoscopy evaluation was randomised using a set of Latin squares. Duplicate videos were randomly interspersed in the video set, but positioned so that they were separated by at least eight other videos; videos comprising a CFT+/− pair were separated by at least four other videos and the viewing order balanced. The order of descriptors was randomised between investigators using Latin squares so that each descriptor appeared first (second, third, etc) an equivalent number of times across investigators, although the order was constant for each investigator. Video clips illustrating each descriptor and anchor points on the Likert scale were provided and data (descriptors on four- or five-point Likert scales,

Table 1 Descriptors and definitions

Descriptor (score most severe lesions)	Likert scale anchor points	Definition
Vascular pattern	Normal (1)	Normal vascular pattern with arborisation of capillaries clearly defined
	Patchy loss (3)	Patchy loss or blurring of vascular pattern
	Obliterated (5)	Complete loss of vascular pattern
Mucosal erythema	None (1)	The colour of the mucosa is normal
	Light red (3)	Some increase in colour of the mucosa that is probably abnormal, but would be best compared side by side with a normal examination
	Dark red (5)	Red or crimson colour of the mucosa that is similar to blood—that is, clearly abnormal even if not compared with a normal examination (does not include intramucosal haemorrhage)
Mucosal surface (Granularity)	Normal (1)	Smooth mucosa with a sharp light reflex, similar to a polished surface
	Granular (3)	Mucosal surface diffuses reflected light causing minor variation in the surface
	Nodular (5)	Evident nodular variation in mucosal surface
Mucosal oedema	None (1)	Normal appearance: no white or yellow substance visible
	Probable (3)	Slight swelling and thickening of mucosa
	Definite (5)	Marked thickening and oedema of the mucosa with blunting of the mucosal folds
Mucopus	None (1)	Normal appearance: no white or yellow substance visible
	Some (3)	White or yellow deposits on the mucosa unrelated to any bowel preparation
	Lots (5)	Mucopus substantially covering the mucosal surface unrelated to any bowel preparation
Bleeding	None (1)	No visible blood
	Mucosal (2)	Some spots or streaks of coagulated blood on the surface of the mucosa ahead of the scope, which can be washed away
	Luminal mild (3)	Some free liquid blood in the lumen
	Luminal moderate (4)	Frank blood in lumen ahead of endoscope or visible oozing from mucosa after washing intraluminal blood
	Luminal severe (5)	Frank blood in the same lumen with visible oozing from a haemorrhagic mucosa
Incidental friability	None (1)	No bleeding or intramucosal haemorrhage before or after passage of the endoscope
	Mild (2)	No bleeding at the site of assessment before, but minor bleeding or intramucosal haemorrhage after, passage of the endoscope
	Moderate (3)	Intramucosal haemorrhage without overt bleeding before passage of the endoscope
	Severe (4)	Overt bleeding after passage of the endoscope
	Very severe (5)	Overt bleeding from the mucosa
Contact friability	None (1)	No bleeding from the mucosa after light touch with closed biopsy forceps
	Probable (3)	Intramucosal haemorrhage or minor bleeding after light touch with closed biopsy forceps
	Definite (5)	Overt bleeding mucosa after light touch (within 10 s) with closed biopsy forceps
Erosions and ulcers	None (1)	Normal mucosa, no visible erosions or ulcers
	Erosions (2)	Tiny ($\leq$5 mm) defects in the mucosa, of a white or yellow colour with a flat edge
	Superficial ulcer (3)	Larger ($>$5 mm) defects in the mucosa, which are discrete fibrin-covered ulcers in comparison with erosions, but remain superficial
	Deep ulcer (4)	Deeper excavated defects in the mucosa, with a slightly raised edge
Extent of erosions or ulcers	None (1)	None seen during endoscopy
	Limited (2)	$<$10% of the affected mucosa
	Substantial (3)	10%—30% of the affected mucosa
	Extensive (4)	$>$30% of the affected mucosa

*An additional descriptor attempted to describe the transition from abnormal to normal mucosa, but was discarded during phase 1 on the basis that it defied definition. Erosions and ulcers had four (response) levels while the others had five because the expert panel were unable to form a range of five responses with meaningful or measurable distinctions between 2 and 3 or 3 and 4.

with overall assessment of severity by VAS) were collected electronically using a programmed PalmPilot. The range of endoscopic severity was graphically checked by plotting the mean severity level evaluated by VAS as a function of its rank order.

Statistics

Intraobserver variation was assessed by κ statistics[14] calculated from the two pairs of duplicate videos. Interobserver variation was stratified by investigator pairs for the common videos they

Table 2 Distribution and allocation of videos to investigators

	Mayo Clinic stratum								
Expected severity stratum	Normal	0	1—2	3—5	6—7	8—9	10—11	Most severe	Total videos
Core videos	5	4	6	8	8	8	6	5	50
Core videos assigned to each investigator	2*	2	2	3	3	3	2	2	19
Duplicates of core video assigned to investigators	—	—	Each investigator was assigned two videos that duplicated two core videos from among these strata					—	2
Contact friability videos (One with CFT, one without CFT)	—	—	2	2	2	2	2	—	10
CFT videos assigned to each investigator	—	—	Each investigator was assigned two CFT pairs, where the CFT+ videos were nominally in these strata.					—	4
Total readings assigned to each investigator	2*	2	2—4	3—5	3—5	3—5	2—4	2	25

One of the videos in the normal stratum was later found to be from a patient, thus there were truly four screening colonoscopies in this stratum.
*Owing to a video error in this stratum, five readers viewed one instead of two normal videos.
CFT, Contact Friability Test.

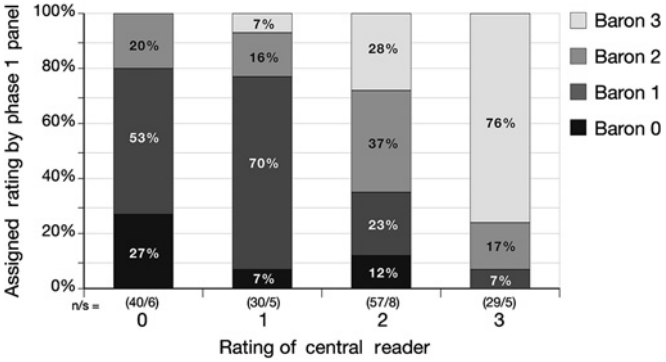

Figure 1 Distribution of levels of Baron score among specialists in the phase 1 panel as a function of the level assigned by the central reader. Ten authors of this paper scored the severity of ulcerative colitis according to their standard practice in 16 videos selected randomly from a total of 24. A level (rating) of the Baron score was then assigned, based on their assessment of friability and this was compared with the level assigned by a central reader. (0= normal; 1=minor; 2=moderate; 3=severe endoscopic severity). n, total number of ratings by phase 1 panel; s, number of video sigmoidoscopies.

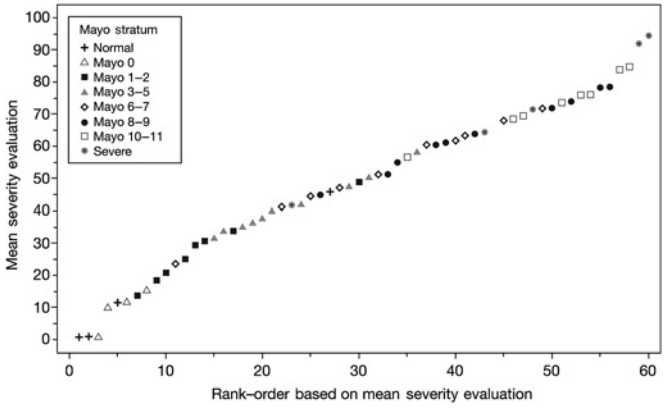

Figure 2 Mean assessment of overall severity as a function of its rank among all mean evaluations of severity, based on 750 evaluations performed by 30 investigators on 25 out of 60 videos. Mean overall severity on a visual analogue scale ranged from 0.67 (video in the normal stratum) to 96.4 (in the most severe stratum) across 25 out of 60 videos scored by 30 investigators, indicating that the videos selected provided an appropriate range of endoscopic severity.

scored, but excluding the second scoring of duplicate and CFT videos, and incomplete data (<5%). An average of investigator-pair κ values ('overall κ') was calculated, where the weighting was the inverse of their variance. Two κ values were calculated: the standard κ summarising the exact level of agreement and a weighted κ taking into account partial credit for disagreement, by assigning a weight of 1 for agreeing levels, 0.5 for levels in adjacent categories on the Likert scale except for the two lowest levels and 0 for any other level. Qualitative interpretation of κ statistics used the convention of Landis and Koch.[15]

Relationships between descriptors and overall severity (VAS) were explored using generalised linear mixed model (GLM) regression. GLM regression used the β distribution for variance and the logit link.[16] The logit link function constrains real parameters to a value between 0 and 1. Descriptors were included as categorical variables, so that the contribution of each level for each descriptor could be explored separately and up to three-way interactions between descriptors assessed. The investigator was included as a random effect. The suitability of models was assessed by plotting least-squares means, examining residual plots and the Akaike Information Criterion (AIC).[17] Described roughly, the AIC is the log likelihood penalised by the number of parameters, a large negative value indicating a good fit in a parsimonious model. An R^2 statistic, denoted pR^2, was the squared correlation between the logit-transformed overall severity evaluations on VAS and linear function of predictors from the model. p Values for tests of specific terms (eg, interactions) were determined from asymptotic F tests. The strategy for optimising the number of response levels was to start with the full number of levels for each item and use the regression modelling (specifically AIC and patterns of mean responses) to identify opportunities to eliminate or combine levels while still maintaining a strong correspondence to the overall score (VAS). All statistical analyses were conducted using SAS V.9.2.

RESULTS

Phase 1

Severity ratings by the 10 IBD specialists showed substantial variation when compared with each other (data not shown), while Baron scores derived from their findings did not match those assigned by a central reader (figure 1). There was 76% agreement for 'severe' activity, but only 27% agreement for a normal appearance and 37% for moderate severity among the videos selected.[6] Ten descriptors (table 1) and full-length

Table 3 Intra-investigator variation results

	Response (%)					κ	
Descriptor	**1**	**2**	**3**	**4**	**5**	**Standard**	**Weighted**
Vascular pattern	3.3	5.0	23.3	11.7	56.7	0.51	0.61
Mucosal erythema	5.0	15.8	39.2	15.8	24.2	0.37	0.43
Mucosal surface	11.7	12.5	35.0	8.3	32.5	0.37	0.45
Mucosal oedema	7.8	11.2	34.5	10.3	36.2	0.33	0.43
Mucopus	30.0	17.5	33.3	8.3	10.3	0.38	0.47
Bleeding	33.3	38.3	15.0	10.0	3.3	0.51	0.57
Incidental friability	24.4	38.3	14.8	15.7	7.0	0.37	0.49
Contact friability (CFT)	23.5	10.8	30.4	8.8	26.5	0.33	0.34
Erosions and ulcers	26.7	32.5	31.7	9.2	—	0.56	0.65
Extent of erosions and ulcers	26.7	32.8	25.9	14.7	—	0.51	0.60

Based on 60 repeat pair assessments (two pairs per investigator) of 36 separate videos with Mayo Clinic scores between 1 and 11. 'Response' for each descriptor refers to the percentage of responses across all assessments. 'Descriptor' refers to the descriptive term used for endoscopic assessment (table 1). Columns 1–5 represent levels on the Likert scale of severity for each item. Erosions and ulcers and extent of erosion and ulcers items had four response levels on the Likert scale; all other items had five levels.
CFT, Contact Friability Test.

Table 4 Interinvestigator variation results

Descriptor	Response (%)					κ	
	1	2	3	4	5	Standard	Weighted
Vascular pattern	11.7	6.8	21.4	8.6	51.4	0.34	0.42
Mucosal erythema	15.6	11.1	36.5	15.7	21.1	0.25	0.35
Mucosal surface	18.9	12.5	31.6	11.7	25.2	0.26	0.34
Mucosal oedema	16.5	12.3	25.7	12.8	32.7	0.23	0.31
Mucopus	37.8	13.3	27.6	8.7	12.5	0.32	0.40
Bleeding	41.9	29.7	14.8	9.0	4.6	0.29	0.37
Incidental friability	30.2	31.5	21.8	9.7	6.9	0.30	0.40
Contact friability (CFT)	25.0	12.8	29.6	7.8	24.7	0.23	0.30
Erosions and ulcers	37.1	27.1	24.8	11.0	—	0.36	0.45
Extent of erosions and ulcers	36.2	21.9	21.3	20.6	—	0.32	0.42

Based on a total of 630 assessments of 60 videos: 21 per investigator with 19 core videos (15 representing Mayo Clinic strata 0—11, two to three per stratum, 2 normal, 2 severe) and two CFT+ videos (representing Mayo Clinic strata 1—11). 'Response' for each descriptor refers to the percentage of responses across all assessments. 'Descriptor' refers to the descriptive term used for endoscopic assessment (table 1). Columns 1—5 represent levels on the Likert scale of severity for each item. Erosions and ulcers and extent of erosion and ulcers items had four response levels on the Likert scale; all other items had five levels.
CFT, Contact Friability Test.

recordings were selected for phase 2. The descriptor discarded after phase 1 was that which attempted to describe the transition from abnormal to normal mucosa, on the basis that it defied definition. Short-length videos were excluded, because of variation in scoring from full-length videos (data not shown) and the risk of editing out information from the original.

Phase 2

Seven hundred and fifty evaluations were performed on 60 videos by 30 investigators (response rate 100% for overall assessment of severity by VAS and ≥96.5% for all descriptors). Mean overall assessments of endoscopic severity scores ranged from a VAS of 0.67 (video in the normal stratum) to 96.4 (most severe stratum), suggesting that the 60 videos encompassed the range of endoscopic severity seen in clinical practice (figure 2).

Intraobserver and interobserver agreement

Sixty repeat pair assessments (two pairs per investigator) of 36 separate videos were assessed for intraobserver variability (table 3). Weighted intrainvestigator κ statistics ranged from 0.34 for contact friability to 0.65 for erosions and ulcers. Six hundred and thirty assessments of 60 videos (21 per investigator, excluding duplicates and CFT-) assessed interobserver variability.

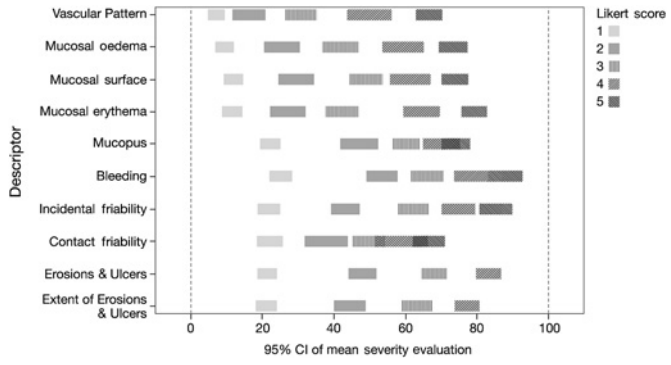

Figure 3 Predicted mean overall assessment of severity for each level of each descriptor. Assessment of overall severity using a 100 point visual analogue scale for each level on the Likert scale of severity for each descriptor (table 1). Predictors are based on generalised linear mixed modelling, using logit link, β distribution for variance, investigator as a random effect and descriptors one by one as categorical variables.

Interinvestigator κ statistic ranged from 0.30 for contact friability to 0.45 for erosions and ulcers (table 4). It is notable that CFT, designed to reduce variation, showed the lowest level of agreement.

Overall assessment of severity

The mean (95% CI) overall assessment of severity according to the 100-point VAS for each descriptor and each level of the Likert scale derived from the GLM model are shown in figure 3. Some descriptors (eg, vascular pattern) appear to provide discrimination for lower levels of severity, with others (eg, bleeding) discriminating at higher levels of severity.

Table 5 Generalised linear mixed models with one, two and three descriptors

Number of descriptors	Descriptors used (number of levels)	AIC	pR^2
1	Erosions and ulcers (4)		
	Mucosal erythema (5)	−607	0.57
	Mucosal oedema (5)	−582	0.55
	Vascular pattern (5)	−561	0.57
	Incidental friability (5)	−495	0.49
	Bleeding (5)	−437	0.44
2	Erosions and ulcers (4) +		
	Mucosal erythema (5), or	−923	0.75
	Vascular pattern (5), or	−887	0.74
	Incidental friability (5)	−866	0.69
3	Erosions and ulcers (4) +		
	Vascular pattern (5) + incidental friability (5), or	−1108	0.91
	Incidental friability (5) + mucosal surface (5), or	−1045	0.90
	Vascular pattern (5) + bleeding (5)	−1038	0.90
3 (Simplified I)*	Erosions & Ulcers (4) +		
	Vascular pattern (5) + incidental friability (4) or	−1132	0.91
	Vascular pattern (5) + bleeding (4)	−1042	0.90
3 (Simplified II)†	Erosions & Ulcers (4) + vascular pattern (3) + bleeding (4)	−999	0.90

pR^2, the squared correlation between the logit-transformed overall severity evaluations on VAS and linear function of predictors from the model.
*Incidental friability and bleeding descriptors simplified to four levels (levels 4 and 5 combined).
†Vascular pattern simplified to three levels (levels 1, 2 and 3 combined), with incidental friability and bleeding as in*.
AIC, Akaike Information Criterion.

Table 6 UCEIS descriptors and definitions

Descriptor (score most severe lesions)	Likert scale anchor points	Definition
Vascular pattern	Normal (1)	Normal vascular pattern with arborisation of capillaries clearly defined, or with blurring or patchy loss of capillary margins
	Patchy obliteration (2)	Patchy obliteration of vascular pattern
	Obliterated (3)	Complete obliteration of vascular pattern
Bleeding	None (1)	No visible blood
	Mucosal (2)	Some spots or streaks of coagulated blood on the surface of the mucosa ahead of the scope, which can be washed away
	Luminal mild (3)	Some free liquid blood in the lumen
	Luminal moderate or severe (4)	Frank blood in the lumen ahead of endoscope or visible oozing from mucosa after washing intraluminal blood, or visible oozing from a haemorrhagic mucosa
Erosions and ulcers	None (1)	Normal mucosa, no visible erosions or ulcers
	Erosions (2)	Tiny ($\leq$ 5mm) defects in the mucosa, of a white or yellow colour with a flat edge
	Superficial ulcer (3)	Larger (>5 mm) defects in the mucosa, which are discrete fibrin-covered ulcers in comparison with erosions, but remain superficial
	Deep ulcer (4)	Deeper excavated defects in the mucosa, with a slightly raised edge

Additional files indicating the levels of the UCEIS are available online only.

Regression modelling to develop an index

GLM model regression was based on a total of 609 assessments of 60 separate videos, excluding second assessments of repeat video pairs; videos with CFT and assessments from an investigator with a large amount of missing data were extracted. The best regression models using one, two and three descriptors are detailed in table 5 (1, 2 and 3), clearly showing an increasing fit with the number of descriptors (lower AIC and higher pR^2). Analysis of the plots of least-squares means indicated that some levels of incidental friability and bleeding could be combined, leading to improvement in AIC values. The best model had four levels for erosions and ulcers and incidental friability, in combination with five levels for vascular pattern, although the model with four levels for erosions and ulcers and bleeding and five levels for vascular pattern had a similar pR^2 (table 5, 3 (simplified I)). However, reducing the vascular pattern to three levels only resulted in a slight loss of fit, with a slightly lower AIC, but similar pR^2 (table 5, 3 (simplified II)). The simplicity of this model and easier definition of three levels of vascular pattern resulted in the selection of this model.

Model selected to create the index

The selected model consists of three descriptors: erosions and ulcers, bleeding and vascular pattern (table 6). Predicted mean severity levels (and 95% CI) for different combinations of Likert scale levels of the three descriptors are shown in table 7. Relationships between actual mean overall assessments of severity (VAS) and means predicted by the model are shown in figure 4. When individual assessments were compared with predicted values, the pR^2 was 0.78. Since the model assigned a level of overall severity to combinations of responses, there is no single coefficient per descriptor.

DISCUSSION

This study has determined that just three descriptors (vascular pattern, bleeding, erosions and ulcers) are sufficient to create a model accounting for the full range of endoscopic severity associated with UC. The UCEIS accurately predicts overall endoscopic severity judged by a VAS, although this needs to be validated by new investigators.

Phase 1 of the study evaluated variability in endoscopic interpretation among specialists in IBD and established definitions of descriptive terms. Phase 2 defined inter- and intra-observer variation, to construct a model to compare with an overall assessment of endoscopic severity. There was widespread variability among specialists in the assessment of endoscopic severity. Disagreement in phase 1 was greatest for videos categorised as 'normal' or 'moderate', with only 27% agreement for normal appearance and 37% for moderate severity, and at best 76% agreement for 'severe' activity.

Phase 2 involved 30 investigators from Europe, USA and Canada. The sample size was large: for intraobserver variation, 60 repeat pairs of 36 videos were used. For interobserver variation, there were 630 assessments of 60 videos. Assessment design was robust: videos were stratified by clinical severity, allowing for greater variability in the mid-range of severity unknown to investigators, then randomly assigned with a random order for scoring descriptors. Reproducibility of scoring within and between investigators was modest, as expected. Interobserver variation is not synonymous with 'agreement', since the latter is not corrected for chance agreement and correction depends on response distribution. It is possible (perhaps even probable) that the variation was due to sampling error, although this could not be quantified, nor allowed for without a substantial increase in sample size. The order of descriptors was randomised to avoid bias, but this may have increased variation between observers, so the descriptor order will be constant in subsequent validations. κ Values may appear poor, but the level of agreement is typical for clinical evaluation processes. For example, evaluating microscopic disease activity in UC reported a κ statistic of 0.20–0.42, improving to 0.59–0.70 with a pictorial scale.[18]

A notable finding was that contact friability was too variable to be further considered. The test, where closed biopsy forceps were pushed against the mucosa to determine whether bleeding occurred, was an construct designed to standardise assessment of mucosal friability in the ASCEND 3 clinical trial,[6] similar to brushing the mucosa with a cotton wool pledget.[3] 'Incidental friability', bleeding from the mucosa seen during withdrawal of the flexible endoscope, was more reproducible. The concept of mucosal friability, however, is poorly understood and always needs explanation. It evaluates mucosal fragility, assumed to be a feature of inflammation before ulceration, where bleeding occurs after minor pressure on the mucosa.

The index (UCEIS) was developed from different combinations of descriptors predicting the overall assessment of severity judged by the investigator using a VAS. Regression techniques established the simplest combination of descriptors most accurately predicting the overall level of severity. Individual descriptors were included as categorical variables, so that each score for each descriptor could be explored separately, including

Table 7 Predicted mean severity index and potential UCEIS grade according to different combinations of Likert scale levels of each of the three descriptors

Erosions and ulcers	Bleeding	Vascular pattern	Predicted severity on a scale 0—100 (95% CI)	Erosions and ulcers	Bleeding	Vascular pattern	Predicted severity on a scale 0—100 (95% CI)
1	1	1	4 (3 to 6)	3	1	1	39 (17 to 67)
1	1	2	18 (15 to 21)	3	1	2	44 (34 to 55)
1	1	3	28 (24 to 34)	3	1	3	60 (53 to 65)
1	2	1	9 (4 to 20)	3	2	1	52 (26 to 77)
1	2	2	29 (24 to 35)	3	2	2	56 (49 to 63)
1	2	3	45 (37 to 53)	3	2	3	65 (60 to 70)
1	3	1	21 (7 to 49)	3	3	1	*
1	3	2	41 (32 to 51)	3	3	2	64 (53 to 73)
1	3	3	56 (44 to 67)	3	3	3	73 (68 to 77)
1	4	1	*	3	4	1	*
1	4	2	54 (38 to 69)	3	4	2	59 (43 to 74)
1	4	3	67 (39 to 86)	3	4	3	80 (75 to 84)
2	1	1	8 (2 to 31)	4	1	1	52 (25 to 77)
2	1	2	25 (21 to 30)	4	1	2	61 (41 to 79)
2	1	3	49 (42 to 56)	4	1	3	73 (63 to 81)
2	2	1	35 (19 to 56)	4	2	1	*
2	2	2	41 (35 to 47)	4	2	2	75 (60 to 86)
2	2	3	54 (49 to 59)	4	2	3	80 (74 to 85)
2	3	1	33 (17 to 54)	4	3	1	*
2	3	2	46 (34 to 59)	4	3	2	*
2	3	3	63 (56 to 69)	4	3	3	78 (68 to 86)
2	4	1	*	4	4	1	*
2	4	2	69 (58 to 79)	4	4	2	92 (79 to 97)
2	4	3	78 (72 to 83)	4	4	3	93 (91 to 95)

The least severe combination (1 each for erosions and ulcers, bleeding and vascular pattern) predicts an index of 4 (95% CI 3 to 6), while the most severe (4 for erosions and ulcers and bleeding, 3 for vascular pattern), predicts an index of 93 (95% CI 91 to 95) on the visual analogue scale (0—100).
*A combination of responses neither seen in the study nor predicted, since they are clinically implausible.

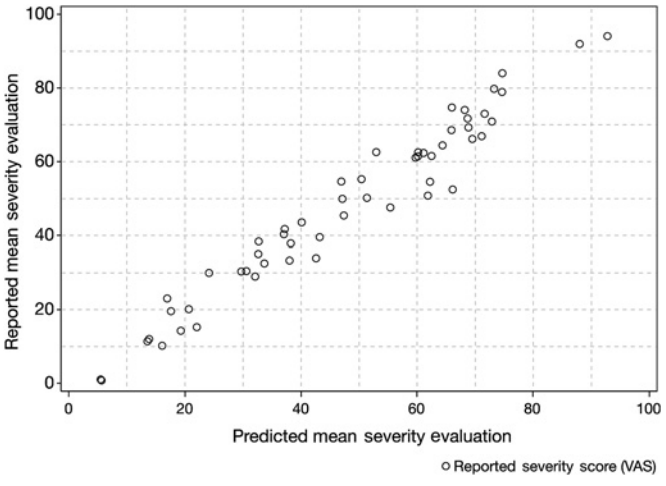

Figure 4 Predicted mean assessment of severity compared with reported mean assessment of severity. To construct the index after excluding the second assessment of repeat video pairs and the videos with a Contact Friability Test (CFT), each of the 30 investigators evaluated 21 independent videos, leading to 630 evaluations. Each video was scored by 10—12 investigators, except for Mayo Clinic score 0 videos, which were scored by 15 investigators (making up the 630). Twenty-one evaluations with missing data were excluded from the index construction (making 609 evaluations overall). Thus, for each video, evaluations by 10—15 investigators were available, allowing the mean of the evaluations of overall severity to be calculated, as well as the mean of the severity evaluations predicted from the generalised linear mixed model using the three descriptors—vascular pattern, bleeding and erosions and ulcers—according to the levels of these predictors reported by each investigator. VAS, visual analogue scale.

interactions between descriptors. One- and two-descriptor models captured 55—75% of the variability in the overall evaluation (table 5). However, several three-descriptor models captured 90—91% of variability, which is a high level of predictability for overall severity assessment. All three-descriptor models included erosions and ulcers. Plots of least-squares means showed that levels on the Likert scale for two of the descriptors (incidental friability and bleeding) could be compressed (from five to four levels) without loss of predictability. Compression of the Likert scale for vascular pattern (to three levels) resulted in some loss of fit, but a pragmatic definition of a fourth level of vascular pattern was impracticable, so this was accepted. This left two leading three-descriptor models, including incidental friability, or bleeding. The latter captured 90% of the variability (pR^2=0.90) and the former 91% (pR^2=0.91), so the choice could not be made on statistical grounds alone. The panel reconvened and decided to include bleeding on the basis of clinical relevance and simplicity.

The terms vascular pattern and bleeding are of course included in the Baron index. Where the UCEIS differs is to define different levels for each of three descriptors, to exclude friability and to apply precise definitions. In theory there are 48 (4×4×3) possible response combinations to the three items. The final index can only assign a value to a fraction of combinations, since some will not be seen in practice and others will be combined after statistical analysis. Validation of potential grades is in progress, but remission might be defined as level 1 for all three descriptors (allowing blurring or loss of capillary margins with a recognisable vascular pattern, no visible bleeding and no erosions or ulceration). On the other hand, 'severe disease' might be defined as a level of at least 3 for vascular pattern and bleeding, with 2 for erosions and ulcers. Such an approach is

likely to bring consistency to endoscopic evaluation of severity, but it is premature to define thresholds.

The 'gold standard' for assessing disease activity in UC should be a diagnostic test that can accurately predict future disease outcome, to augment clinical evaluation. Endoscopy is a surrogate end point and it needs to be established that the UCEIS correlates with, and predicts, clinical outcome. Future studies should test (head to head) whether this instrument can predict clinical outcome better than clinical assessment (without endoscopy) or biomarkers (eg, faecal calprotectin or lactoferrin). The burden of proof has to be on endoscopy, as an expensive and invasive test, to prove that it is better than non-invasive and less expensive alternatives.

A new index for disease activity in ulcerative colitis (the UCEIS) has been created. It illustrates that there is wide variation in the endoscopic interpretation of disease severity between observers. Just three descriptors can be combined to account for 90% of the overall assessment of endoscopic severity judged by a VAS. The UCEIS is undergoing independent validation with different videos and investigators, evaluating operating properties of the index (responsiveness and reliability). Minimal differences for this instrument remain to be evaluated, for its role in research, training and practice.

Author affiliations
[1]Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK
[2]Middletown, Ohio, USA
[3]Warner Chilcott, Weybridge, UK
[4]Division of Gastroenterology, University of Miami Leonard M. Miller School of Medicine, Miami, Florida, USA
[5]Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[6]Hôpital Claude Huriez, Centre Hospitalier Universitaire de Lille, Lille, France
[7]Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, Ontario, Canada
[8]Section of Gastroenterology & Nutrition, University of Chicago Medical Center, Chicago, Illinois, USA
[9]Service de Gastroentérologie, Université Paris Diderot, Hôpital St Louis, Paris, France
[10]Division of Gastroenterology, Department of Medicine, University of Pennsylvania, Philadelphia, USA
[11]AP-HP, Hôpital Lariboisière Medicosurgical Department of Digestive Diseases and University Denis Diderot, Paris, France
[12]Universitätsklinik Innere Medizin III, Abteilung Gastroenterologie und Hepatologie, Medical University of Vienna, Vienna, Austria
[13]Mount Sinai Hospital, New York City, New York, USA
[14]Division of Digestive Diseases, University of Cincinnati, Cincinnati, Ohio, USA
[15]Bernhardt Regulatory Consulting, Cincinnati, Ohio, USA
[16]INSERM U717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France
[17]Division of Gastroenterology, University of California San Diego, La Jolla, California, USA

Acknowledgements We sadly acknowledge the untimely death of Marc Lémann, one of the co-authors of this study who made unparalleled contributions to this and to so many other areas of gastroenterology. Biostatistical advice was both independent and conducted by the sponsors of the study (Procter and Gamble Pharmaceuticals, later Warner Chilcott), although it was established from the outset that the index would be freely available subject to copyright, but not to patent. We are particularly grateful to the investigators who evaluated video endoscopies in phase 2, from Austria (Walter Reinisch, Vienna); Belarus (Yury Marakhouski); Canada (Robert Bailey, Edmonton; Marc Bradette and Gilles Jobin, Quebec; Naoki Chiba, Guelph, Flavio Habal, Toronto; John Marshall, Hamilton); Croatia (Davor Stimac, Rijeka); Estonia (Riina Salupere, Tartu); Hungary (György Székely, Budapest); Italy (Silvio Danese, Milan); Latvia (Juris Pokrotnieks, Riga); Poland (Jaroslaw Regula, Warsaw); Romania (Mircea Manuc, Bucharest); Russia (Olga Alexeeva, Nizhegorodskiy); Serbia (Njegica Jojic, Belgrade) and the USA (Nelson Ferreira, Hagerstown, MD; Fred Fowler, Harrisburg, NC; Daniel Geenen, Milwaukee, WI; Norman Gilinsky, Cincinatti, OH; Howard Gus, Ocean, NJ; Asher Kornbluth, New York, NY; Mark Lamet, Hollywood, FL; Jacque Noel, Lafayette, LA; Michael Safdi, Cincinnati, OH; Jerrold Schwartz, Arlington Heights, IL; Guarang Shah, Jacksonville, FL; Larry Weprin, Dayton, OH; Estephan Zayat, Wichita, OH). We would also like to acknowledge Barry Rodgers-Gray for assistance with the figures, Mr Scott Hayes (Procter and Gamble), who provided the data acquisition and data management support for the study and Professor Bryan Warren, Oxford, who originally suggested using the endoscopic videos from a randomised controlled trial in this way.

ICMJE disclosures have been submitted.

Funding Procter and Gamble, later Warner Chilcott. Sponsors paid for video image preparation, managerial support, statistical programming and for the time spent by independent gastroenterologists to evaluate the UCEIS. All authors gave freely of their time and have received no remuneration for the development of this index.

Competing interests None.

Ethics approval Ethics approval was provided by ASCEND: EUDRACT 2006-001310-32 and Oxford LREC 536407Q1605/58ORH.

Contributors The manuscript was written by SPLT, PK, JYM and WJS. DS performed the satistical analyses, with further evaluation by JYM and DGA. MTA, JFC, BGF, SBH, ML, GRL, PRM, BES, WJS and SPLT evaluated the videos in phase 1 and WR in phase 2. CAB and PK coordinated the planning and implementation, supported by BRY. All authors except ML critically appraised the manuscript.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Supplementary files and original data are available on request.

REFERENCES

1. **Cooney RM,** Warren BF, Altman DG, *et al*. Outcome measurement in clinical trials for ulcerative colitis: toward standardisation. *Trials* 2007;**8**:17—25.
2. **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—86.
3. **Baron JH,** Connell AM, Lennard-Jones JE. Variation between observers in describing mucosal appearances in proctocolitis. *Br Med J* 1964;**1**:89—92.
4. **Sutherland LR,** Martin F, Greer S, *et al*. 5-Aminosalicylic acid enema in the treatment of distal ulcerative colitis, proctosigmoiditis, and proctitis. *Gastroenterology* 1987;**92**:1894—8.
5. **Kamm MA,** Sandborn WJ, Gassull M, *et al*. Once-daily, high-concentration MMX mesalamine in active ulcerative colitis. *Gastroenterology* 2007;**132**:66—75.
6. **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—43, e1—3.
7. **Travis S,** Cooney R, Lukas M, *et al*. Conduct of clinical trials in UC: impact of independent scoring of endoscopic severity on results of a randomised controlled trial with a peptide and 5-ASA. *Am J Gastroenterol* 2006;**101**(suppl 9):S429.
8. **Orlandi F,** Brunelli E, Feliciangeli G, *et al*. Observer agreement in endoscopic assessment of ulcerative colitis. *Ital J Gastroenterol Hepatol* 1998;**30**:539—41.
9. **de Lange T,** Larsen S, Abaakken L. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. *BMC Gastroenterol* 2004;**4**:9.
10. **Travis SP,** Higgins PD, Orchard T, *et al*. Review article: defining remission in ulcerative colitis. *Aliment Pharmacol Ther* 2011;**34**:113—24.
11. **de Lange T,** Larsen S, Abaakken L. Image documentation of endoscopic findings in ulcerative colitis: photographs or video clips? *Gastrointest Endosc* 2005;**61**:715—20.
12. **Gomes P,** duBoulay CD, Smith CL, *et al*. Relationship between disease activity indices and colonoscopic findings in patients with colonic inflammatory bowel disease. *Gut* 1986;**27**:92—5.
13. **Pera A,** Bellando P, Caldera D, *et al*. Colonoscopy in inflammatory bowel disease. Diagnostic accuracy and proposal of an endoscopic score. *Gastroenterology* 1987;**92**:181—5.
14. **Chmura Kraemer H,** Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002;**21**:2109—29.
15. **Landis JR,** Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;**33**:363—74.
16. **Ferrari S,** Cribari-Neto F. Beta regression modeling for rates and proportions. *J Appl Statist* 2004;**31**:799—815.
17. **Akaike H.** A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;**AC-19**:716—23. (http://garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf).
18. **Geboes K,** Riddell R, Ost A, *et al*. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut* 2000;**47**:404—9.