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

Travis SPL, Schnell D, Krzeski P, Abreu MT, Altman DG, Colombel JF, Feagan BG, Hanauer SB, Lémann M, Lichtenstein GR, Marteau PR, Reinisch W, Sands BE, Yacyshyn BR, Bernhardt CA, Mary JY, Sandborn WJ.

CORRESPONDENCE TO:

Dr Simon P L Travis.

Address: Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, OX3 9DU, UK.

Telephone: +44 1865 228753

Fax: +44 1865 228763

Email: simon.travis@ndm.ox.ac.uk

Word count: Abstract 300; Text 3368 (reduced from 3845);

Abbreviations:

AIC: Akaike Information Criterion; CFT: Contact Friability Test; GLM: generalised linear mixed model regression; IBD: inflammatory bowel disease; UC: ulcerative colitis; UCEIS: Ulcerative Colitis Endoscopic Index of Severity; VAS: Visual Analogue Scale;

Key words: ulcerative colitis, endoscopy, disease severity, activity index, instrument development

CO-AUTHORS:

Maria T Abreu

Division of Gastroenterology, University of Miami Leonard M. Miller School of Medicine, Miami, Florida, USA.

Douglas G Altman

Centre for Statistics in Medicine, University of Oxford, Oxford, UK.

Christian A Bernhardt

Bernhardt Regulatory Consulting, Cincinnati, Ohio USA

Jean-Frédéric Colombel

Hôpital Claude Huriez, Centre Hospitalier Universitaire de Lille, Lille, France.

Brian G Feagan

Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, Ontario, Canada.

Stephen B Hanauer

University of Chicago, USA.

Marc Lémann

Service de Gastroentérologie, Université Paris Diderot, Hôpital St Louis, Paris, France.

Gary R Lichtenstein

Division of Gastroenterology, Department of Medicine, University of Pennsylvania, Philadelphia, USA.

Jean–Yves Mary

INSERM U717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France.

Phillippe R Marteau

AP-HP, Hôpital Lariboisière Medicosurgical Department of Digestive Diseases & University Denis Diderot, Paris 7, France.

Piotr Krzeski

Warner Chilcott, Weybridge, UK

Walter Reinisch

Universitätsklinik Innere Medizin III, Abteilung Gastroenterologie und Hepatologie, Medical University of Vienna, Vienna, Austria.

William J Sandborn

Division of Gastroenterology, University of California San Diego, La Jolla, California, USA.

Bruce E Sands

Mount Sinai Hospital, New York City, NY, USA.

Dan Schnell

Middletown, ~~Procter and Gamble, Mason,~~ Ohio, USA.

Bruce R Yacyshyn

Division of Digestive Diseases, University of Cincinnati, Cincinnati, OH, USA.

Word count: Abstract 298; Text 3845;

References 18

Tables 7;

Figures 4

Key words: ulcerative colitis, endoscopy, disease severity, activity index, instrument development

ABSTRACT

Objective

Variability in endoscopic assessment necessitates rigorous investigation of descriptors for scoring ~~the~~ severity ~~of~~ of ulcerative colitis (UC). ~~Our aims were~~We aimed to evaluate variation in the overall endoscopic assessment of severity, the intra- and inter-individual variation of descriptive terms, and to create an ulcerative colitis endoscopic index of severity (UCEIS) which could be validated.

Design

~~We performed a~~A two-phase study us~~ed~~ing a library of 670 videosigmoidoscopies from patients with Mayo Clinic scores 0-11, supplemented by 10 videos from 5 people without UC and 5 hospitalised patients with acute severe UC. In Phase 1, each of ten investigators viewed 16/24 videos to assess agreement on Baron score with a central reader and agreed definitions of 10 endoscopic descriptors. In Phase 2, each of 30 ~~other~~ different investigators rated 25/60 different videos for the ~~10~~ descriptors ~~on 4-5 level Likert scales~~ and assessed overall ~~assessment of~~ severity on a 0-100 visual analogue scale. Kappa statistics tested inter- and intra-observer variability for each descriptor. A general linear mixed regression model based on logit link and beta-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 P~~p~~hase I investigators and the central reader. In Phase 2, weighted kappas ranged from 0.34-0.65 and 0.30-0.45 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 & Ulcers (none/erosions/superficial/deep) ~~graded by 3-4 level Likert scales~~, each with precise definitions, which explained 90% of the variance (pR^2, Akaike Information Criterion -999) of 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 UCEIS accurately predicts overall assessment of endoscopic severity of UC. Validity and responsiveness need further testing before ~~this instrument~~it can be applied as an outcome measure in clinical trials or ~~in~~ clinical practice.

What is already known about this subject?

- There is wide variation in the endoscopic assessment of severity of ulcerative colitis
- There is no validated instrument ~~for endoscopic assessment~~

What are the new findings

- An index (UCEIS) ~~has been developed~~ with three descriptors~~ive terms~~ (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 3 or 4 levels with precise definitions
- Friability is excluded from the endoscopic description of severity

How might this impact on clinical practice

- Once independently validated, the UCEIS will be available for clinical trials, training and practice

INTRODUCTION

Endoscopy plays a pivotal role in the evaluation of ulcerative colitis (UC). At least 9 different scoring systems are used as outcome measures in clinical trials and endoscopy plays an important role in most [1,2]. I~~The i~~ndices are typically composite measures that include assessment of symptom severity, quality of life, laboratory tests and endoscopic findings. However, the ~~net~~ contribution of endoscopy ~~-~~is index-specific~~-~~. In the widely used Mayo Clinic index [2], endoscopy~~ic findings~~ is~~are~~ one of four criteria and ~~, while they are one of~~ just one of two criteria (in addition to rectal bleeding) currently used by the Food and Drug Administration (FDA) for defining remission. Consequently, inter-observer variation in ~~the~~ assessing ~~ment of~~ endoscopic activity is important, because disagreement ~~between observers~~ can alter the proportion of patients defined as in remission and influence regulatory decisions~~conclusions~~.

The original endoscopic grading of UC (Baron index, 1964) was developed ~~in 1964, long~~ before index methodology ~~and criteria for index development~~ w~~e~~as~~re~~ defined [3]. It has~~-~~ been used, nevertheless, in most trials of active UC to this day, with only minor and unvalidated modification [2]. ~~Nevertheless, d~~Data supporting the index are scant. It was created by scoring 7 endoscopic descriptors in 60 patients, ~~made~~ by 3 observers using rigid sigmoidoscopes. The kappa statistic, a measure of inter-observer 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 ~~of activity~~ [3] by using the presence of mucosal friability to discriminate between mild (Baron level 1) and moderately active (Baron level 2) disease [4~~-~~6~~,5~~]. ~~Friability represents the ease with which the mucosa can be damaged and made to bleed, so a contact friability test was developed in an attempt to standardise the degree of damage [6].~~

Variation between observers in categorising endoscopic disease activity is widely suspected [1,2,7-10], so the need for this to be quantified ~~y this~~ appears self-evident. The aims of this study were first to substantiate variation in endoscopic assessment of activity in UC, ~~and~~ then to evaluate intra- and inter-

individual variation of descriptive terms, and finally to create an ulcerative colitis endoscopic index of severity (UCEIS) which could be validated.

METHODS

The study ~~was conducted in~~included two phases. Phase 1 ~~was designed to~~ mapped inconsistencies in endoscopic assessment and defined the most dependable descriptive terms ('descriptors'). Phase 2 ~~was designed to characterise and~~ quantified~~y~~ inter- and intra-observer 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 ~~used to~~for assessing activity; 'descriptor' refers to an ~~individual~~ item within that index with severity allocated on a Likert scale; and ~~the word~~ 'level' is used to refer to the severity graded for an item. ~~The term~~ 'Sscore' 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~~ies~~ according to their own practice, ~~but~~ in the absence of clinical information. 24 representative videos were selected to represent the widest range of ulcerative colitis activity, guided by the Mayo Clinic score (by PK & 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. Sufficient quality recordings (sharp image, sufficient bowel preparation) were selected. Videos from fibreoptic endoscopes were discarded. 16 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 inter-observer variation. Each investigator was randomly assigned 16 of the 24 videos in randomized order using a set of Latin Squares: ~~.~~ a core set of 8 videos that all investigators evaluated (2

per Baron score) and 8 of 16 non-core videos, This kept the number of evaluations by each investigator in the 2-3 hour 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 length of video [11], each investigator had two pairs that were ~~full-length/~~shortened from 10-15 minutes to approximately 5 minutes, 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). ~~A rating for each assessment was subsequently derived by applying the Baron index [3] to the findings recorded by the investigator. 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 substantiate variability in endoscopic ~~interpretation~~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 3-, 4- or 5-point Likert scales of severity for each descriptor, were subsequently agreed by consensus during a videoteleconference between investigators (Table 1).

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 intra-mucosal 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 intra-luminal 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 endoscope
	Mild (2)	No bleeding at the site of assessment before, but minor bleeding or intramucosal haemorrhage after the passage of endoscope

	Moderate (3)	Intramucosal haemorrhage without overt bleeding before the 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 sec) with closed biopsy forceps
Erosions & Ulcers	None (1)	Normal mucosa, no visible erosions or ulcers
	Erosions (2)	Tiny (≤5mm) defects in the mucosa, of a white or yellow colour with a flat edge
	Superficial ulcer (3)	Larger (>5mm) defects in the mucosa, which are discrete fibrin-covered ulcers when compared to 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 observed during endoscopy
	Limited (2)	Less than 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 & Ulcers had 4 (response) levels while the others had 5 because the expert panel were unable to form a range of 5 responses with meaningful or measureable distinctions between 2&3 or 3&4.

Phase 2

Fifty core videos were assembled, 40 from the library of 670 videos (by PK & BRY, representing Mayo Clinic levels (scores) 0-11, different from those selected for phase 1), representing 6 expected severity strata (note selection criteria for Phase 1). A further 5 from individuals without UC and 5 from patients hospitalised with severe UC who had consented to their anonymised images being used for study

(Oxford LREC 536407Q1605/58ORH), represented 2 additional strata at the expected extremes of endoscopic severity. The 5 patients admitted with biologically severe UC represented the most severe end of the spectrum of UC, although only 2/5 came to colectomy (one within 6 hours of flexible sigmoidoscopy). To evaluate the Contact Friability Test (CFT), 10 different videos ~~from the same library~~ representing Mayo Clinic levels 1-11, 2 per 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 US and Canada (see acknowledgements) scored 25 videos selected from the 60 recordings, but blinded to clinical information or Mayo Clinic ~~level~~score. Endoscopists were recruited to reflect a range of geographic and institutional characteristics from investigators with endoscopic training in trials of inflammatory bowel disease (IBD) or known to authors with an interest in endoscopy or IBD (840 median colonoscopies and flexible sigmoidoscopies/yr (range 100-2100), median 25 years' endoscopy experience , range 8-35). E~~As described in Table 2, e~~ach investigator was randomly allocated a CDROM containing 15 out of 40 core videos from the library (comprising 2 to 3 videos selected from each of 6 Mayo Clinic levels), 2 out of 5 normal videos from people without UC and 2 out of 5 videos from patients with severe UC, together with 2 out of 10 CFT+/- pairs (Table 2). To evaluate intra-observer consistency, each investigator also scored 2 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 ~~visual analogue scale (~~VAS~~)~~.

Table 2: Distribution and allocation of videos to investigators

Expected severity stratum	**Mayo Clinic Stratum**								Total videos
	Normal	0	1-2	3-5	6-7	8-9	10-11	Most severe	
Core videos	5*	4	6	8	8	8	6	5	50
Core videos assigned to each	2**	2	2	3	3	3	2	2	19

investigator									
Duplicates of core video assigned to investigators	-	-	Each investigator was assigned 2 videos that duplicated 2 core videos from among these strata					-	2
Contact Friability videos (1 with CFT, 1 without CFT)	-	-	2	2	2	2	2	-	10
CFT videos assigned to each investigator	-	-	Each investigator was assigned 2 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 4 screening colonoscopies in this stratum.

**Due to a video error in this stratum, 5 readers viewed 1 instead of 2 Normal videos.

Videos were assigned to investigators using an incomplete block design, stratified by expected severity stratum. T~~he result of t~~his randomisation process meant~~was~~ that each video was scored by 10-12 investigators, except for the 4 videos in ~~the~~ Mayo Clinic level 0 stratum, which were each viewed by 15 investigators. Due to an assignment error, 5/30 investigators were ~~only~~ assigned only ~~one~~ 1 and not ~~two~~ 2 normal videos. The order of endoscopy evaluation was randomized using a set of Latin Squares. Duplicate videos were randomly interspersed in the video set, but positioned s~~ouch~~ that they were separated by at least 8 other videos; videos comprising a CFT+/- pair were separated by at least 4 other videos and the viewing order balanced. The order of descriptors was randomized 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 4 or 5 point Likert scales, with overall assessment of severity by VAS) were collected electronically using a programmed Palm Pilot®.

~~A large number of v~~The~~ideos were selected that could reasonably be expected to cover the whole~~ range of endoscopic severity ~~, which~~ was graphically checked by plotting the mean severity level evaluated by VAS as a function of its rank order. ~~Investigators were recruited to reflect a range of geographic and institutional characteristics. All patients (including those outside clinical trials) gave permission for their anonymised videos to be independently reviewed.~~

Statistics

Intra-observer variation was assessed by kappa statistics [14] calculated from the two pairs of duplicate videos. Inter-observer variation was ~~assessed by kappa statistics [14],~~ stratified by investigator pairs for the common videos they scored, but excluding the second scoring of duplicate and CFT- videos, as well as incomplete data (<5%). An average of investigator-pair kappas ('overall kappa') was calculated, where the weighting was the inverse of their variance. ~~For both intra- and inter-observer variation, t~~Two kappas were calculated: the standard kappa summarising the exact level of agreement and a weighted kappa taking into account partial credit for disagreement, by assigning a weight of 1 for agreeing levels, ~~of~~ 0.5 for levels in adjacent categories on the Likert scale except for the 2 lowest levels, and ~~of~~ 0 for any other level. Qualitative interpretation of kappa statistics used the convention of Landis and Koch [15].

Relationships between descriptors and overall severity (~~scored by~~ VAS) were explored using generalised linear mixed model (GLM) regression. GLM regression used the beta distribution for variance and the logit link [16]. The logit link function constrains ~~the~~ real parameters to a value between 0 and 1. Descriptors were included ~~in the models~~ as categorical variables, so that the contribution of each level for each descriptor could be explored separately and up to 3-way interactions between descriptors ~~were~~ assessed. The investigator was included as a random effect. The suitability of models was assessed by plotting ~~the~~ least-squares' means, examining residual plots and ~~by~~ the Akaike Information Criterion (AIC) [17]. Described roughly, the AIC is the log likelihood penali~~z~~sed by the

number of parameters ~~in the model~~, a large negative value indicating a good fit in a parsimonious model. An R^2 statistic, denoted ~~as~~ pR^2, was ~~defined as~~ 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 (e~~.~~g~~.~~ interactions) ~~in models~~ 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 modeling (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 version 9.2.

RESULTS

Phase 1

S~~Overall s~~everity ratings by the 10 IBD specialists showed substantial variation when compared with each other (data not shown), while ~~the~~ Baron scores derived from their findings did not match ~~the Baron score~~those assigned by a central reader (Figure 1). There was 76% agreement for 'severe' activity, ~~and 70% for minor,~~ 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 recordings were selected for Phase 2 ~~of the study~~. The descriptor discarded after phase I was an attempt 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

750 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 (~~video in~~

~~the~~ most severe stratum) ~~-~~suggesting that the 60 ~~selected~~ videos encompassed the range of endoscopic severity seen in clinical practice (Figure 2).

Intra- and inter-observer agreement

60 repeat pair assessments (2 pairs per investigator) of 36 separate videos were assessed for intra-observer variability (Table 3). W~~The w~~eighted intra-investigator kappa statistics ranged from 0.34 for Contact Friability to 0.65 for Erosions & Ulcers. 630 ~~. Six hundred and thirty~~ assessments of 60 videos (21 per investigator, excluding duplicates and C~~R~~FT-) ~~were used to~~ assessed inter-observer variability. I~~The i~~nter-investigator kappa statistic ranged from 0.30 for Contact Friability to 0.45 for Erosions & Ulcers (Table 4). It is notable that CFT, ~~which was~~ designed to reduce ~~inter-observer~~ variation, ~~did not~~ demonstrated the lowest level of ~~a level of~~ agreement ~~higher than most other descriptors~~.

Table 3: Intra-investigator variation results

Descriptor	Response (%)					Kappa	
	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 & Ulcers	26.7	32.5	31.7	9.2	-	0.56	0.65
Extent of Erosions & Ulcers	26.7	32.8	25.9	14.7	-	0.51	0.60

Legend**:** Based on 60 repeat pair assessments (2 pair 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 & Ulcers and Extent of Erosion & Ulcers items had 4 response levels on the Likert scale; all other items had 5 levels.

Table 4: Inter-investigator variation results

Descriptor	**Response (%)**					**Kappa**	
	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 & Ulcers	37.1	27.1	24.8	11.0	-	0.36	0.45
Extent of Erosions & Ulcers	36.2	21.9	21.3	20.6	-	0.32	0.42

Legend**:** Based on a total of 630 assessments of 60 videos: 21 per investigator with 19 core videos (15 representing Mayo Clinic strata 0-11, 2 to 3 per stratum, 2 normal, 2 severe) and 2 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 & Ulcers and Extent of Erosion & Ulcers items had 4 response levels on the Likert scale; all other items had 5 levels.

Overall assessment of severity

The mean (~~and its~~ 95% confidence interval (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 (e~~.~~g~~.~~ Vascular Pattern) appear to provide ~~more~~ discrimination for lower levels of severity, w~~h~~il~~st~~h others (e~~.~~g~~.~~ Bleeding) ~~appear to~~ discriminatinge~~d~~ at higher levels of severity.

Regression modelling to develop an index

GLM model regression was based on a total of 609 assessments of 60 separate videos, excluding ~~the~~ second assessments of repeat~~ed~~ video pairs~~:~~, ~~the~~ videos with ~~the~~ CFT ~~extracted~~ and ~~the~~ 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 4 levels for Erosions & Ulcers and Incidental Friability, in combination with 5 levels for Vascular Pattern, although the model with 4 levels for Erosions & Ulcers and Bleeding, and 5 levels for Vascular Pattern had a similar pR^2 (Table 5, 3 (simplified I)). However, reducing the Vascular Pattern to 3 levels only resulted in 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 3 levels of Vascular Pattern resulted in the selection of this model. ~~Best models with one, two or three descriptors are presented in Table 5 with AIC, pR^2 values and number of levels for each descriptor.~~

~~*Single descriptor models*. Individual descriptors that correlated best with the overall assessment of severity by VAS were Mucosal Erythema and Mucosal Oedema (Table 5, 1 descriptor).~~

~~*Two-descriptor models.* The best regression models that used two descriptors (and their interactions) were those with Erosions & Ulcers combined with Mucosal Erythema, Vascular Pattern, or Incidental Friability Table 5, 2 descriptors). Plotting the least squares' means of different descriptors showed overlap in the confidence limits of some levels of the Likert scales, meaning that some levels could be combined (figures not shown). There was greater separation between levels for the combination of Bleeding and Incidental Friability than for Mucosal Erythema and Vascular Pattern (asymptotic F-test: $p<0.001$), which also captured a similar range of overall severity (from 10 to 92 vs 12 to 93, respectively). This indicated that there may be value of adding bleeding or friability to a model combining the Erosions & Ulcers descriptor with Vascular Pattern or Mucosal Erythema.~~

~~*Three-descriptor models*. The top 3 models are shown in table 5, for up to 3 descriptors. By reducing the number of levels on Likert scales for Incidental Friability and Bleeding descriptors, AIC values improved. The top model had 4-level Likert scales for Erosions & Ulcers and Incidental Friability in combination with 5 levels for Vascular Pattern, although the model with 4 levels for Erosions & Ulcers and Bleeding, and 5 levels for Vascular Pattern had a similar pR^2 (Table 5, 3 (simplified I)). Reducing the Vascular Pattern to a 4-level descriptor resulted in small changes in the AIC, with slight loss of fit when further simplified to 3 response levels (Table 5, 3 (simplified II)).~~

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

Number of descriptors	**Descriptors used** (number of levels)	**AIC**	**pR^2**
1	Erosions & 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 & Ulcers (4) +		
	Mucosal Erythema (5), or	- 923	0.75
	Vascular Pattern (5), or	- 887	0.74
	Incidental Friability (5)	- 866	0.69
3	Erosions & 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

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

Selected model to create the index

The ~~final~~ selected model ~~selected~~ consists of three descriptors: Erosions & Ulcers, Bleeding, and Vascular Pattern (~~with the last 2 descriptors being on 4 and 3 point scales, respectively, rather than the original 5-point scales,~~ Table 6). Predicted mean severity levels (and 95% confidence interval) for different combinations of Likert scale levels of the three descriptors are shown in table 7. Relationships between ~~the~~ actual mean overall assessments of severity (VAS) and means predicted by th~~e~~is model are shown ~~(in~~ Figure 4~~)~~. When individual assessments were compared to predicted values, the pR^2 was 0.78. Since the

models directly assigned a level of overall severity to combinations of responses, there is no single coefficient per descriptor.

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 intra-luminal blood, or visible oozing from a haemorrhagic mucosa
Erosions & Ulcers	None (1)	Normal mucosa, no visible erosions or ulcers
	Erosions (2)	Tiny (≤5mm) defects in the mucosa, of a white or yellow colour with a flat edge
	Superficial ulcer (3)	Larger (>5mm) defects in the mucosa, which are discrete fibrin-covered ulcers when compared to erosions, but remain superficial
	Deep ulcer (4)	Deeper excavated defects in the mucosa, with a slightly raised edge

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 & Ulcers	**Bleeding**	**Vascular Pattern**	**Predicted severity on a scale 0-100** (95% CI)	**Erosions & Ulcers**	**Bleeding**	**Vascular Pattern**	**Predicted severity on a scale 0-100** (95% CI)
1	1	1	4 (3-6)	3	1	1	39 (17-67)

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

Legend: The least severe combination (1 each for Erosions & Ulcers, Bleeding and Vascular pattern) predicts an index of 4 (95% CI 3-6), while the most severe (4 for Erosions & Ulcers and Bleeding, 3 for Vascular pattern), predicts an index of 93 (95% CI 91-95) on the visual analogue scale (0-100).

†: a combination of responses neither observed in the study nor predicted, since they are clinically implausible

DISCUSSION

~~In spite of limited reproducibility of the endoscopic desciptors selected for evaluation, this study has allowed a new ulcerative colitis index of severity (UCEIS) to be constructed.~~ Thise study has determined that just three descriptors (Vascular Pattern, Bleeding, Erosions & Ulcerss~~, Bleeding and Vascular Pattern~~) a ~~we~~re sufficient to create a model ~~that~~ accountinged for the full range of endoscopic severity associated with UC ~~(from normal mucosa to severe colitis preceding colectomy)~~. ~~Our results show that t~~The UCEIS accurately predicts overall endoscopic severity judged by a visual analogue scale, although~~. However,~~ this ~~finding~~ needs to be validated by new investigators ~~before it can be applied as an outcome measure in clinical trials, used for training, or influence treatment decisions in practice~~.

P

~~The approach involved two phases: the first~~ hase 1 of the study evaluated ~~the~~ variability in endoscopic interpretation among specialists in IBD and established definitions of descriptive terms.~~;~~ Phase 2 ~~the second~~ defined inter- and intra-observer variation ~~among another set of investigators~~, ~~so that a model could be~~ to construct a model ~~ed~~ to compare with an overall assessment of endoscopic severity. ~~A large resource of 670 videos recorded in a clinical trial of active UC recorded according to a standard protocol was used. Specialists in Phase 1 simply represented a group of investigators interested in clinical trials of UC. Phase 1 demonstrated the existence of~~There was widespread variability among specialists in the assessment of endoscopic severity. Disagreement in Phase 1 was greatest for videos categorised ~~by the central reader~~ as 'normal' or 'moderate, with only 27% agreement for ~~a~~ normal appearance and 37% for

moderate severity among the videos selected, although there wasand at best only 76% agreement for 'severe' activity.

Phase 2 involved a wide range of 30 investigators from Europe, the USA and Canada, based on contributions to other clinical trials of UC, or recommendations by the two lead investigators. The sample size was large: to assessfor intra-observer variation, 60 repeat pairs of 36 videos were used. F, while for inter-observer variation, there were 630 assessments of 60 videos. AThe assessment design was robust: videos were first stratified by clinical severity, with an allowance made for probableallowing for increased greater variability in the mid-range of severity unknown to investigators, then randomly assigned with a random order for scoring descriptors. Reproducibility of scoring by the same individualwithin and between investigators was modest, as anticipated. It should be noted that defining inter-observer variation is not synonymous with 'agreement', since the latter is not corrected for chance agreement and the correction depends on response distribution. It is possible, (perhaps even probable) ,that a large amount of the variation we observed was due to sampling error, although this could not readily be quantified, nor could it be allowed for without a substantial increase in sample size. The order of the descriptors might also have contributed. Although tThe order of descriptors was randomised to avoid bias, but this random order of descriptors may have increased variation between observers, so we will use the same descriptor order will be constant across investigators in subsequent validation. The Kappa values may appear poor, but the level of agreement for individual descriptors is nevertheless typical for clinical evaluation processes. For example, a study evaluating variation between histopathologists scoring microscopic disease activity in ulcerative colitisUC found reported a n initial kappa statistic of 0.20-0.42, that improvinged to 0.59-0.70 with a pictorial scale for each component [18]. A notable finding from analysing inter-observer variability was that Contact Friability was too variable to be further considered. The Contact Friability Testtest, where closed biopsy forceps were pushed against the mucosa to determine whether bleeding occurred, had beenwas an artificial construct designed to standardise the assessment of mucosal friability in the ASCEND 3 clinical trial [6], similar to

brushing the mucosa with a cotton wool pledget ~~described by Baron et al.~~ [3]. 'Incidental Friability', ~~describing~~ bleeding from the mucosa seen during withdrawal of the flexible endoscope, ~~turned out to be~~was more reproducible. The concept of mucosal friability, however, is poorly understood and always needs ~~to be~~ explanation~~ined to endoscopists~~. It evaluates ~~the~~ mucosal fragility ~~of the mucosa~~, assumed to be a feature of inflammation before ulceration, where~~by~~ bleeding occurs after minor pressure on ~~degrees of contact with~~ the mucosa.

The index (UCEIS) was ~~then~~ developed ~~by examining the ability of~~from different combinations of descriptors ~~to~~ predicting the overall assessment of severity ~~as~~ judged by the investigator using a visual analogue scale. ~~Our goal was to use~~ R~~r~~egression techniques ~~to find~~ established the simplest combination of descriptors ~~that~~ most accurately predicting~~ed~~ the overall level of severity. Individual descriptors were included ~~in the model~~ as categorical variables, so that ~~the contribution of~~ each score for each descriptor could be explored separately, including interactions between descriptors. One- and two-descriptor models ~~only~~ captured 55-75% of the variability in the overall evaluation ~~of severity~~ (table 5). However, several three-descriptor models captured 90-91% of variability, which is a high level of predictability for ~~the~~ overall severity assessment. All ~~these~~ three-descriptor models included Erosions & Ulcers~~, which had the highest intra- and inter-observer agreement~~. Plots of ~~the~~ least squares' means showed that ~~lthe number of l~~evels on the Likert scale for two of the descriptors (incidental friability and bleeding) could be compressed (from 5 to 4 levels) without loss of predictability. Compression of the Likert scale for Vascular Pattern ~~(~~to ~~(~~3 levels) resulted in some loss of fit, but ~~it was not possible to find~~ a pragmatic definition of a fourth level of Vascular Pattern was impracticable, so ~~this loss of predictability~~this was accepted. This left two leading three-descriptor models, ~~one~~ including ~~Erosions & Ulcers, Vascular Pattern and~~ Incidental Friability, or ~~. The other included Erosions & Ulcers, Vascular Pattern and~~ Bleeding. The latter ~~was as sensitive~~captured 90% of the variability (pR^2=0.90~~, capturing 90% of the~~

variability in overall severity) as the former (pR^2=0.91), so t. The choice between including Incidental Friability or Bleeding in the final model could not be made on statistical grounds alone. The panel therefore reconvened and decided to include Bleeding on the grounds of clinical relevance and simplicity.

The terms vascular pattern and bleeding are of course included in the Baron index which (for instance) includes vascular pattern in its description of normal mucosa and either spontaneous bleeding or bleeding to light touch (friability) for more severe endoscopic activity. Ulceration was not included in the Baron index, because no ulcers were seen in the 60 patients examined. Where the UCEIS differs is to define different levels for each of three descriptors, to exclude friability and to apply precise definitions. The success of using different combinations of these terms is shown by its ability to represent the full range of endoscopic severity from normal to 'worst case', representing the mucosa in hospitalised patients prior to colectomy. Potential interactions between descriptors mean that evaluating a simple scoring system whereby points are assigned for each descriptor and then summed (i.e. a 'main effects' model) may not be statistically sound. 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 these combinations, since some combinations will not be observed in practice and others will be combined after statistical analysis. Nominal grades for the UCEIS might be assigned to illustrate the potential for discriminating between endoscopic remission, mild, moderate and severe disease. Validation of potential grades is in progress, but it can be seen how remission might be defined as a level 1 for all three descriptors (allowing some blurring or loss of capillary margins with , but a recognisable Vascular Pattern, no visible bleeding and no erosions or ulceration). On the other hand, 'severe disease' might be defined for the purposes of a clinical trial as a level of at least 3 for Vascular Pattern (complete loss) and Bleeding (free blood in the lumen), with 2 for Erosions & Ulcers (tiny ≤5mm flat erosions in the mucosa). Such an approach is likely to bring greater consistency to endoscopic evaluation of severity, but it is premature to define thresholds at this stage. `

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 endpoint 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 novel index for ~~scoring~~ disease activity in ulcerative colitis (the UCEIS) has been created. ~~The current study~~It illustrates the ~~limitations of subjective assessment of complex pictures and~~ confirms that there is wide variation in the endoscopic interpretation of disease severity between observers. Just three descriptors ~~, each with 3 or 4 levels of severity defined by consensus and subjected to analysis by least squares' means,~~ can be combined to account for 90% of the overall assessment of endoscopic severity judged by a visual analogue scale. The UCEIS is ~~now~~ undergoing independent validation with different ~~groups of~~ videos and investigators, evaluating . ~~The~~ operating properties of the index (responsiveness and reliability). ~~will be evaluated in future trials of therapy. Further work will be needed to define t~~Mhe minimalum ~~clinically important~~ differences for this instrument remain t~~and t~~o be evaluated, for its role in research, training and ~~clinical~~ practice.

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 & 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, ~~and~~ to Mr Scott Hayes (Procter & 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. ~~-~~

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;**5375:**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:**1894-98.

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.

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 SPL**, Higgins PDR, Orchard T, Van der Woude CJ, Panacionne R, Bitton A, O'Morain C, Panès J, Sturm A, Reinisch W, Kamm MA, D'Haens G. Review article: Defining remission in ulcerative colitis. *Aliment Pharmacol Ther* 2011 (in press).

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:**925.

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. **Kraemer HC**, 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 modelling for rates and proportions. *J Appl Statist* 2004;**31:**799-815.

17. **Akaike H.** A new look at the statistical model identification. *IEEE Transaction on Automatic Control* 1974;**AC-19:**716-23.

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.

Figure legends

Figure 1: 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

Legend: 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.

Figure 2: Distribution of levels of Baron score among specialists in the phase 1 panel as a function of the level assigned by the central reader

Legend: 10 authors of this paper scored the severity of UC according to their standard practice in 16 videos selected randomly out of 24 videos. 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 endoscpic severity). n = total number of ratings by phase 1 panel; s = number of video-sigmoidoscopies.

.

Figure 3: Predicted mean overall assessment of severity for each level of each descriptor

Legend: 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, beta distribution for variance, investigator as a random effect and descriptors one by one as categorical variables.

Figure 4: Predicted mean assessment of severity compared to reported mean assessment of severity

Legend: To construct the index after excluding the second assessment of repeat video pairs and the videos with CFT, each of the 30 investigators evaluated 21 independent videos, leading to 630 evaluations. Each video was scored by 10 to 12 investigators, except for Mayo Clinic score 0 videos, which were scored by 15 investigators (making up the 630). 21 evaluations with missing data were excluded from the index construction (making 609 evaluations overall). Thus, for each video, evaluations by 10 to 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 GLM model using the 3 descriptors Vascular Pattern, Bleeding and Erosions & Ulcers, according to the levels of these predictors reported by each investigator. ~~Reported mean severity evaluations are the mean investigator evaluations for the 60 videos used in regression modelling. The predicted mean severity evaluation is from generalised linear mixed modelling, using logit link, beta distribution for variance, investigator as a random effect and three descriptors as categorical variables: Erosions & Ulcers (4 levels), Bleeding (4 levels), and Vascular Pattern (3 levels).~~

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non-exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd and its Licensees to permit this article (if accepted) to be published in Gut and any other BMJPGL products to exploit all subsidiary rights, as set out in our licence (http://group.bmj.com/products/journals/instructions-for-authors/licence-forms).