**Original Article**

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


**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. 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 1 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 (pR², 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.
in most trials of active UC to this day, with only minor and unvalidated modification. Data supporting the index are scant. It was created by scoring seven endoscopic descriptors in 60 patients, by three observers using rigid sigmoidoscopes. The k 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 by using the presence of mucosal friability to discriminate between mild (Baron level 1) and moderately active (Baron level 2) disease.\textsuperscript{3–6}

Variation between observers in categorising endoscopic disease activity is widely suspected,\textsuperscript{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\textsuperscript{10} (EUDRACT 2006-001510-32). 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,\textsuperscript{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.\textsuperscript{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.\textsuperscript{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 556407/Q1605/580RH), represented two additional strata at the expected extremes of endoscopic severity. The five patients admitted with severely 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 50 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 15 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,
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 score severity level evaluated by VAS as a function of its rank order.

Statistics
Intraobserver variation was assessed by \( \kappa \) statistics calculated from the two pairs of duplicate videos. Interobserver variation was stratified by investigator pairs for the common videos they

### Table 1 Descriptors and definitions

<table>
<thead>
<tr>
<th>Descriptor (score most severe lesions)</th>
<th>Likert scale anchor points</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vascular pattern</td>
<td>Normal (1)</td>
<td>Normal vascular pattern with arborisation of capillaries clearly defined</td>
</tr>
<tr>
<td></td>
<td>Patchy loss (3)</td>
<td>Patchy loss or blurring of vascular pattern</td>
</tr>
<tr>
<td>Obliterated (5)</td>
<td></td>
<td>Complete loss of vascular pattern</td>
</tr>
<tr>
<td>Mucosal erythema</td>
<td>None (1)</td>
<td>The colour of the mucosa is normal</td>
</tr>
<tr>
<td>Light red (3)</td>
<td></td>
<td>Some increase in colour of the mucosa that is probably abnormal, but would be best compared side by side with a normal examination</td>
</tr>
<tr>
<td>Dark red (5)</td>
<td></td>
<td>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)</td>
</tr>
<tr>
<td>Mucosal surface (Granularity)</td>
<td>Normal (1)</td>
<td>Smooth mucosa with a sharp light reflex, similar to a polished surface</td>
</tr>
<tr>
<td>Granular (3)</td>
<td></td>
<td>Mucosal surface diffuses reflected light causing minor variation in the surface</td>
</tr>
<tr>
<td>Nodular (5)</td>
<td></td>
<td>Evident nodular variation in mucosal surface</td>
</tr>
<tr>
<td>Mucosal oedema</td>
<td>None (1)</td>
<td>Normal appearance: no white or yellow substance visible</td>
</tr>
<tr>
<td>Probable (3)</td>
<td></td>
<td>Slight swelling and thickening of mucosa</td>
</tr>
<tr>
<td>Definite (5)</td>
<td></td>
<td>Marked thickening and oedema of the mucosa with blunting of the mucosal folds</td>
</tr>
<tr>
<td>Mucopus</td>
<td>None (1)</td>
<td>Normal appearance: no white or yellow substance visible</td>
</tr>
<tr>
<td>Some (3)</td>
<td></td>
<td>White or yellow deposits on the mucosa unrelated to any bowel preparation</td>
</tr>
<tr>
<td>Lots (5)</td>
<td></td>
<td>Mucopus substantially covering the mucosal surface unrelated to any bowel preparation</td>
</tr>
<tr>
<td>Bleeding</td>
<td>None (1)</td>
<td>No visible blood</td>
</tr>
<tr>
<td>Mucosal (2)</td>
<td></td>
<td>Some spots or streaks of coagulated blood on the surface of the mucosa ahead of the scope, which can be washed away</td>
</tr>
<tr>
<td>Luminal mild (3)</td>
<td></td>
<td>Some free liquid blood in the lumen</td>
</tr>
<tr>
<td>Luminal moderate (4)</td>
<td></td>
<td>Frank blood in lumen ahead of endoscope or visible oozing from mucosa after washing intraluminal blood</td>
</tr>
<tr>
<td>Luminal severe (5)</td>
<td></td>
<td>Frank blood in the same lumen with visible oozing from a haemorrhagic mucosa</td>
</tr>
<tr>
<td>Incidental friability</td>
<td>None (1)</td>
<td>No bleeding or intramucosal haemorrhage before or after passage of the endoscope</td>
</tr>
<tr>
<td>Mild (2)</td>
<td></td>
<td>No bleeding at the site of assessment before, but minor bleeding or intramucosal haemorrhage after, passage of the endoscope</td>
</tr>
<tr>
<td>Severe (4)</td>
<td></td>
<td>Overt bleeding after passage of the endoscope</td>
</tr>
<tr>
<td>Very severe (5)</td>
<td></td>
<td>Overt bleeding from the mucosa</td>
</tr>
<tr>
<td>Contact friability</td>
<td>None (1)</td>
<td>No bleeding from the mucosa after light touch with closed biopsy forceps</td>
</tr>
<tr>
<td>Probable (3)</td>
<td></td>
<td>Intramucosal haemorrhage or minor bleeding after light touch with closed biopsy forceps</td>
</tr>
<tr>
<td>Definite (5)</td>
<td></td>
<td>Overt bleeding mucosa after light touch (within 10 s) with closed biopsy forceps</td>
</tr>
<tr>
<td>Erosions and ulcers</td>
<td>None (1)</td>
<td>Normal mucosa, no visible erosions or ulcers</td>
</tr>
<tr>
<td>Erosions (2)</td>
<td></td>
<td>Tiny ((&lt;5) mm) defects in the mucosa, of a white or yellow colour with a flat edge</td>
</tr>
<tr>
<td>Superficial ulcer (3)</td>
<td></td>
<td>Larger ((5) mm) defects in the mucosa, which are discrete fibrin-covered ulcers in comparison with erosions, but remain superficial</td>
</tr>
<tr>
<td>Deep ulcer (4)</td>
<td></td>
<td>Deeper excavated defects in the mucosa, with a slightly raised edge</td>
</tr>
<tr>
<td>Extent of erosions or ulcers</td>
<td>None (1)</td>
<td>None seen during endoscopy</td>
</tr>
<tr>
<td>Limited (2)</td>
<td></td>
<td>(&lt;10% of the affected mucosa</td>
</tr>
<tr>
<td>Substantial (3)</td>
<td></td>
<td>10%–30% of the affected mucosa</td>
</tr>
<tr>
<td>Extensive (4)</td>
<td></td>
<td>&gt;30% of the affected mucosa</td>
</tr>
</tbody>
</table>

*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.

### Table 2 Distribution and allocation of videos to investigators

<table>
<thead>
<tr>
<th>Expected severity stratum</th>
<th>Mayo Clinic stratum</th>
<th>Core videos</th>
<th>Core videos assigned to each investigator</th>
<th>Duplicates of core video assigned to investigators</th>
<th>Contact friability videos (One with CFT, one without CFT)</th>
<th>CFT videos assigned to each investigator</th>
<th>Total readings assigned to each investigator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Normal</td>
<td>0</td>
<td>1–2</td>
<td>3–5</td>
<td>6–7</td>
<td>8–9</td>
</tr>
<tr>
<td>Core videos</td>
<td></td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Duplicate core video assigned to each investigator</td>
<td>2*</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Duplicates of core video assigned to investigators</td>
<td>-</td>
<td>-</td>
<td>Each investigator was assigned two videos that duplicated two core videos from among these strata</td>
<td>-</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact friability videos</td>
<td>(One with CFT, one without CFT)</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CFT videos assigned to each investigator</td>
<td>-</td>
<td>-</td>
<td>Each investigator was assigned two CFT pairs, where the CFT+ videos were nominally in these strata</td>
<td>-</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total videos assigned to each investigator</td>
<td>2*</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

*Owing to a video error in this stratum, five readers viewed one instead of two normal videos.

CFT, Contact Friability Test.

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.

---

**Inflammatory bowel disease**

Inflammatory bowel disease

Table 3  Intra-investigator variation results

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

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.
recording 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 defined 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.

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.
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 model and easier de

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 pR² 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 intraobserver 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 650 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. χ 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 χ statistic of 0.20–0.42, improving to 0.59–0.70 with a pictorial scale.10

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.7 ‘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...
Predicted mean assessment of severity compared with Figure 4 ulcers Bleeding Erosions and descriptors Table 7 Predicted mean severity index and potential UCEIS grade according to different combinations of Likert scale levels of each of the three was scored by 10 evaluated 21 independent videos, leading to 630 evaluations. Each video with a Contact Friability Test (CFT), each of the 30 investigators excluding the second assessment of repeat video pairs and the videos reported mean assessment of severity. To construct the index after the evaluations of overall severity to be calculated, as well as the mean evaluations by 10 0 videos, which were scored by 15 investigators (making up the 630). 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 ($4^3$) 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
1Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK
2Middletown, Ohio, USA
3Warner Chilcott, Weybridge, UK
4Division of Gastroenterology, University of Miami Leonard M. Miller School of Medicine, Miami, Florida, USA
5Centre for Statistics in Medicine, University of Oxford, Oxford, UK
6Hôpital Claude Huriez, Centre Hospitalier Universitaire de Lille, Lille, France
7Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, Ontario, Canada
8Section of Gastroenterology & Nutrition, University of Chicago Medical Center, Chicago, Illinois, USA
9Service de Gastroentérologie, Université Paris Diderot, Hôpital St Louis, Paris, France
10Division of Gastroenterology, Department of Medicine, University of Pennsylvania, Philadelphia, USA
11AP-HP, Hôpital Lariboisière Medico-surgical Department of Digestive Diseases and University Denis Diderot, Paris, France
12Universitätsklinik Innere Medizin III, Abteilung Gastroenterologie und Hepatologie, Medical University of Vienna, Vienna, Austria
13Mount Sinai Hospital, New York City, New York, USA
14Division of Digestive Diseases, University of Cincinnati, Cincinnati, Ohio, USA
15Bernhardt Regulatory Consulting, Cincinnati, Ohio, USA
16INSERM U717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France
17Division of Gastroenterology, University of California San Diego, La Jolla, California, USA
18Division of Gastroenterology, University of California, Los Angeles, California
19INSERM U717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France
20Clinical Trials, Robarts Research Institute, University of Western Ontario, Ontario, Canada
21Department of Gastroenterology, University of California San Diego, La Jolla, California, USA
22INSEECO, Paris, France
23INSERM U717 Biostatistics and Clinical Epidemiology, Université Paris Diderot, Paris, France
24Division of Gastroenterology, Department of Medicine, University of California, San Diego, CA, USA
25Institute of Clinical and Experimental Medicine, Prague, Czech Republic
26Division of Gastroenterology, University of California, Los Angeles, California, USA
27University of California, San Diego, CA, USA
28New York, NY; Mark Lane, Hollywood, FL; Jacque Noel, Lafayette, LA; Michael Salfd, Cincinnati, OH; Jerrod Schwartz, Arlington Heights, IL; Guang Shang, Jacksonville, FL; Larry Weprin, Dayton, OH; Estephon Zayat, Wichita, OH. We would also like to acknowledge Barry Rodgers-Grey for assistance with the figures, Ml 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 2008-001316-32 and Oxford UREC 538407/0165/5/80/R0H.

Contributors The manuscript was written by SPT, PK, JYM and WJS. DS performed the statistical analyses, with further evaluation by JYM and DGA. MTA, JFC, BGF, SBH, MT, GII, 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
Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS)


Gut 2012 61: 535-542 originally published online October 13, 2011
doi: 10.1136/gutjnl-2011-300486

Updated information and services can be found at:
http://gut.bmj.com/content/61/4/535

These include:

Supplementary Material

Supplementary material can be found at:
http://gut.bmj.com/content/suppl/2011/10/13/gutjnl-2011-300486.DC1

References

This article cites 18 articles, 3 of which you can access for free at:
http://gut.bmj.com/content/61/4/535#BIBL

Open Access

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See: http://creativecommons.org/licenses/by-nc/2.0/ and http://creativecommons.org/licenses/by-nc/2.0/legalcode.

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections

Articles on similar topics can be found in the following collections

Open access (350)
Ulcerative colitis (1113)
Colon cancer (1547)
Endoscopy (1003)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/