Journal of Crohn's and Colitis, 2018, 794–803
doi:10.1093/ecco-jcc/jjy035
Advance Access publication March 28, 2018
Original Article

Original Article

Development of Reliable, Valid and Responsive Scoring Systems for Endoscopy and Histology in Animal Models for Inflammatory Bowel Disease

Pim J. Koelink[a], Manon E. Wildenberg[a,b], Larry W. Stitt[c], Brian G. Feagan[c,d,e], Martin Koldijk[f], Angélique B. van 't Wout[f], Raja Atreya[g], Michael Vieth[h], Johannan F. Brandse[a,b], Suzanne Duijst[a], Anje A. te Velde[a], Geert R. A. M. D'Haens[b,c], Barrett G. Levesque[c,i], Gijs R. van den Brink[a,b,j]

[a]Tytgat Institute for Liver and Intestinal Research, Academic Medical Center Amsterdam, The Netherlands [b]Department of Gastroenterology and Hepatology, Academic Medical Center, Amsterdam, The Netherlands [c]Robarts Clinical Trials Inc., London, Ontario, Canada and Amsterdam, The Netherlands [d]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada [e]Department of Epidemiology & Biostatistics, University of Western Ontario, London, Ontario, Canada [f]Janssen Prevention Center of Janssen Vaccines & Prevention BV, Janssen Pharmaceutical Companies of Johnson & Johnson, Leiden, the Netherlands [g]Department of Medicine 1, University of Erlangen-Nürnberg, Erlangen, Germany [h]Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany [i]Department of Medicine, Division of Gastroenterology, University of California San Diego, La Jolla, California, USA [j]Current address: GlaxoSmithKline, Medicines Research Center, Stevenage, UK.

Corresponding author: Pim J. Koelink, Tytgat Institute for Liver and Intestinal Research, Academic Medical Center, Meibergdreef 68–70, Amsterdam 1105 BK, The Netherlands. Tel: +31-20-566-8160; Fax: +31-20-566-9190; E-mail: p.j.koelink@amc.nl

Abstract

Background and Aims: Although several endoscopic and histopathologic indices are available for evaluating the severity of inflammation in mouse models of colitis, the reliability of these scoring instruments is unknown. Our aim was to evaluate the reliability of the individual items in the existing indices and develop new scoring systems by selection of the most reliable index items.
Methods: Two observers scored the histological slides [$n = 224$] and endoscopy videos [$n = 201$] from treated and untreated Interleukin[IL]-10 knock-out andT-cell transferred SCID mice. Intra-rater and inter-rater reliability for endoscopy and histology scores, and each individual item, were measured using intraclass correlation coefficients [ICCs]. The Mouse Colitis Histology Index [MCHI] and Mouse Colitis Endoscopy Index [MCEI] were developed using the most reliable items. Both were correlated to the colon density and to each other and were evaluated for their ability to detect changes in pathobiology.
Results: The intraclass correlation coefficients (ICCs) for inter-rater agreement (95% CIs) for the total histology and endoscopy scores were 0.90 [0.87–0.92] and 0.80 [0.76–0.84], respectively. The MCHI and MCEI were highly correlated with colon density, with a Spearman Rho = 0.81[0.75–0.85] and 0.73 [0.66–0.79], respectively, and with each other, Spearman Rho = 0.71 [0.63–0.77]. The MCHI and MCEI were able to distinguish between the experimental groups within the models, with pairwise differences between the treated and untreated groups being statistically significant [$p < 0.001$].
Conclusions: These histological and endoscopic indices are valid and reliable measures of intestinal inflammation in mice, and they are responsive to treatment effects in pre-clinical studies.

Key Words: Basic science; experimental pathophysiology

1. Introduction

The use of animal studies is critical in addressing basic and translational research questions in inflammatory bowel disease [IBD]. However, a common challenge for both these animal models and patient-based research, is the choice of study end points. Semi-quantitative evaluation of the intestinal histopathology is considered a 'gold standard' for measuring disease severity in animal models of IBD. Such evaluations commonly use ordinal categorical scales, e.g. from normal to severe.[1] Often, ordinal subscores measuring different items, such as immune cell infiltration and crypt loss, are added to obtain a total severity score. It is noteworthy that the relevance of histopathological items may vary between different animal models of IBD.[2] For instance, epithelial destruction is a highly relevant item in the acute dextran sodium sulphate [DSS] model, while epithelial hyperplasia is relevant in the T-cell transfer model. This has led to the introduction of different histopathologic scoring systems for different animal models of IBD. However, even within the same model a variety of scoring systems have been used. The initial scoring system for the T-cell transfer model of colitis, developed by Powrie and Read, used a semi-quantitative grade on a scale from 0 [no change] to 4 [most severe].[3] Over the years, multiple variations of this scoring system have been used in which scores ranged from 0 to 5[4] and even 0 to12.[5] Ostanin et al., who published an article with tips-and-tricks for this T-cell transfer model, used an alternative scoring system [with similar items] that ranged from 0 to 22, but subsequently updated the scoring system with modified scores varying between 0 and 17.[6,7] Similarly, for the Interleukin-10 knock-out [IL10 KO] mouse, another frequently used model for chronic intestinal inflammation, several histopathological scoring systems have been described,[8–11] with items similar to those used for the T-cell transfer model. It is notable that the final disease severity scores are just simple combinations of the item scores, and that none of the overall disease severity indices were developed using a methodologically rigorous scheme of index development,[12] which would have led to the inclusion of only the relevant and reliable items in the final index. This raises important concerns regarding the validity of the disease severity indices that are currently used.

A parallel situation currently exists in the use of endoscopy in murine models of IBD.[13] As is the case with the previously described histopathological indices, existing endoscopic instruments are semi-quantitatively scored and evaluate multiple items.[14,15] Again, as for the histopathologic assessments, multiple variants exist,[16] all of which have been empirically created.

Virtually no research has been conducted for assessing the operating characteristics of the scoring systems used for IBD animal models, in contrast to endoscopic and histopathological scoring systems used in human IBD.[17] Bleich et al., 2004[8] compared the Jackson Laboratory [TJL] and Medical School Hannover [MHH] scores for histopathology, both of which are modifications of the score developed by Berg et al.,[9] with respect to their ability to identify quantitative trait loci [QTL] in IL10 KO mice. They showed that the simplified MHH score had less power to detect QTL compared with the more refined TJL score. The number of categories within an ordinal scoring method indeed has potential implications for the study: a small number can reduce the ability of the scoring system to detect differences between groups, thereby increasing the required number of animals per group. On the other hand, with a large number of categories, readers may not be able to reliably distinguish between the categories. A scoring system evaluating histological or endoscopic severity should be [1] valid [i.e. measures what it is intended to measure], [2] reliable [i.e. is consistently scored] and [3] responsive [i.e. able to detect differences between groups where differences in disease status exist]. These operating properties are mutually exclusive in their scope. For example, an index can be highly reliable but relatively unresponsive to change in disease status.

In clinical IBD, endoscopic and histopathologic indices are routinely used as outcome measures in controlled clinical trials and, to a lesser extent, for patient evaluation in practice. In the past few years, considerable effort has gone into evaluating these instruments, using rigorous methodologies that are also appropriate for the evaluation of the operating properties of the murine indices. Given this background, our objectives were [1] to evaluate the reliability of semi-quantitative scoring systems for endoscopy and histopathology in both the T-cell transfer model and the IL10 KO model, [2] to derive new indices for both endoscopy and histology using reliable items, [3] to assess the responsiveness of the new indices to experimental treatment effects and [4] to evaluate index validity by assessing correlations with colon density, an independent measure of disease severity in animal models of IBD.

2. Materials and Methods

2.1 Animals

All animal studies were approved by the local Animal Ethical Committee and performed according to national guidelines. Female C.B-17 SCID mice and wild-type BALB/c mice were ordered from Harlan [Boxmeer, The Netherlands]. As previously described, colitis was induced by injecting 1–3 × 10^5 $CD4^+CD45RB^{High}$ cells into SCID mice intraperitoneally [i.p.]. Control mice did not receive any cells. Three different experiments were performed. In two experiments, animals were treated with anti-TNFα antibodies, two times per week i.p., from 3 weeks after the T-cell transfer for 2–4 weeks. Female C57BL/6J-$IL10^{tm1Cgn}$ mice [IL10 KO] and wild-type C57BL/6J mice [3–4 weeks of age] were ordered from Jackson Laboratories [Bar Harbor, USA]. Three independent experiments were performed. In

Table 1. Histology scoring items [related to Figure 1].

Score	Inflammatory infiltrate	Goblet cell loss	Crypt density	Crypt hyperplasia	Muscle thickening	Submucosal inflammation	Crypt abcess	Ulceration
0	none	none	normal	none	none	none	absent	absent
1	increased presence of inflammatory cells	<10%	decreased by <10%	slight increase in crypt length	slight	individual cells		
2	infiltrates also in submucosa	10–50%	decreased by 10–50%	2–3-fold increase in crypt length	strong	infiltrate[s]		
3	transmural	>50%	decreased by >50%	>3-fold increase in crypt length	excessive	large infiltrate[s]	present	present

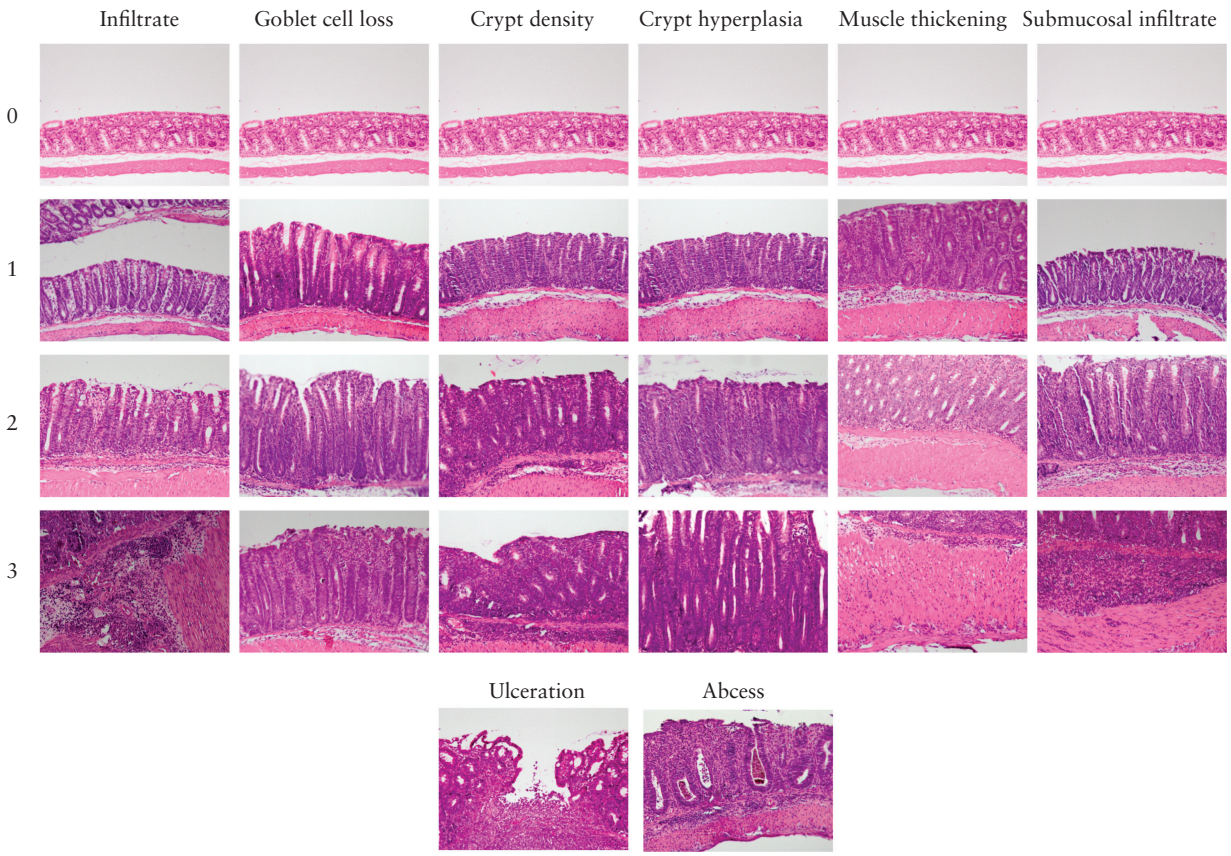

Figure 1. Items in the histology score. Related to Table 1.

two experiments, animals were treated with anti-IL12/23p40 antibodies [300 µg, two times per week i.p] from 8 weeks of age until 14–16 weeks of age. Therapeutic antibodies were a kind gift from Dr D. Shealy [Janssen Research & Development].

2.2 Colon density and histology

Mice were sacrificed by CO_2 suffocation within 24 h after the endoscopic procedure. The large intestine was removed and its length measured when it was in a relaxed position without stretching; it was weighed after removal of the feces. The colon density [mg/cm] was calculated by dividing the weight by the length of the large intestine. The intestine was opened longitudinally and washed thoroughly with phosphate-buffered saline [PBS] then processed as a 'Swiss roll' in paraffin. Five-micrometer slides were cut and stained with hematoxylin and eosin [H&E] using a standard protocol. Slides were scored for tissue quality ['poor' or 'moderate to perfect']. Based on the existing literature, eight histological components were assessed: 'inflammatory infiltrate', 'goblet cell loss', 'hyperplasia', 'crypt density', 'muscle thickness', 'submucosal infiltration', 'ulcerations' and 'crypt abscesses' [all categorized from 0–3, Table 1/Figure 1]. A total histological severity score, ranging from 0 to 24, was obtained by summing the eight item scores.

2.3 Endoscopy

For endoscopic evaluations, mice were anesthetized with 2–3% isoflurane/O_2, and feces were removed as much as possible using flexible feeding tubes [20ga × 30 mm, Instech Laboratories, Inc. Plymouth Meeting, USA]. The Olympus URF type V endoscope was rectally inserted for a maximum of 5 cm, and videos of the endoscopic procedure were recorded using a Medicap USB200 Medical Digital Video Recorder, while retracting the endoscope. Videos were scored for quality ['poor' or 'moderate to perfect']. Five items of endoscopic severity were scored: 'mucosal thickening', 'vasculature', 'granularity of the mucosal surface', 'fibrin deposits' and 'stool appearance' as proposed by Becker et al. [Table 2, Figure 2].[14] The total endoscopic disease severity score was calculated from the items' scores, excluding the stool component score [as it was clearly determinable in only 158 of 201 of the videos], resulting in a total score ranging from 0 to 12.

Table 2. Endoscopy scoring items [related to Figure 2].

Score	Thickening	Vasculature	Fibrin	Granularity	Stool
0	transparent	normal	none	none	normal
1	moderate	moderate	little	moderate	still shaped
2	marked	marked	marked	marked	unshaped
3	non-transparent	absent/bleeding	extreme	extreme	spread

2.4 Study design

For both the histological slides and endoscopy videos, readers were blinded to the experimental group within each model, and videos and images were reviewed in random order by two independent observers. Videos and images were reviewed a second time by the same observers with at least one week in between the first and second reads. The videos and histology slides were also evaluated by two external observers, one for endoscopy and a second for histology, who, other than the item definitions, were not given specific scoring instructions. Global severity of histological or endoscopic

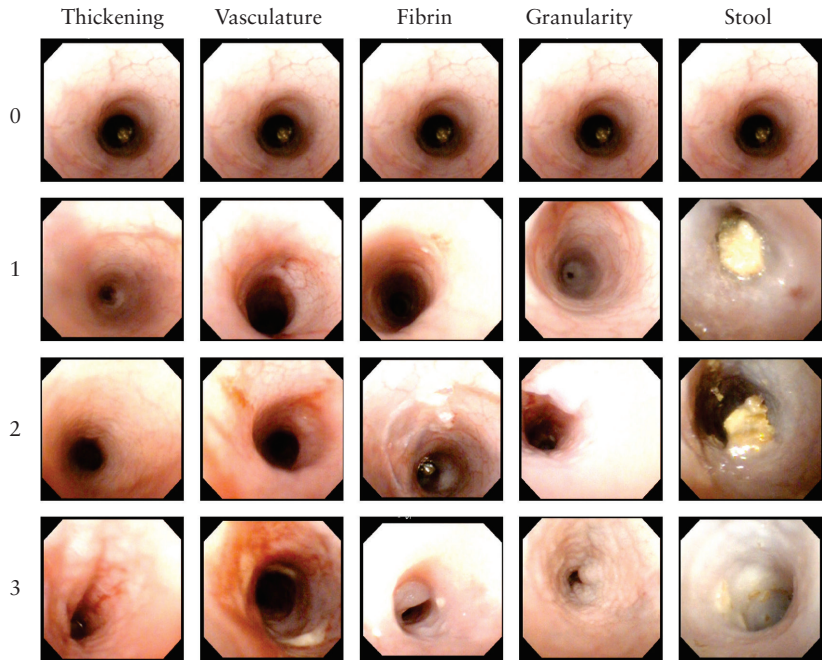

Figure 2. Items in the endoscopy score. Related to Table 2.

Table 3. Animal models and treatment conditions.

Model	Treatment	Dose	Total number [% of total]	Histology [*n*]	Endoscopy [*n*]
T-cell Transfer	Healthy control[a]		31 [22.8%]	30	30
	Placebo		46 [33.8%]	46	42
	Anti-TNFα	1 µg	12 [8.8%]	12	10
	Anti-TNFα	5 µg	12 [8.8%]	12	10
	Anti-TNFα	25 µg	12 [8.8%]	12	12
	Anti-TNFα	100 µg	23 [16.9%]	23	21
	Total		136 [100%]	135	125
IL-10 KO	Healthy control[b]		32 [35.2%]	30	28
	Placebo		35 [38.5%]	35	27
	Anti-IL12/23	300 µg	24 [26.4%]	24	21
	Total		91 [100%]	89	76
Total			227	224	201

[a]SCID mice without a T-cell transfer. [b]Wild-type C57Bl6 mice.

Table 4. Reliability coefficients [95% CI] for the histopathological item scores.

	ICC [95% CI]	
	Inter-rater	Intra-rater
Inflammation	0.76 [0.71–0.80]	0.79 [0.74–0.82]
Goblet cell loss	0.75 [0.68–0.80]	0.89 [0.86–0.92]
Crypt density	0.79 [0.72–0.84]	0.84 [0.79–0.88]
Hyperplasia	0.75 [0.69–0.79]	0.80 [0.75–0.83]
Muscle thickness	0.55 [0.45–0.62]	0.63 [0.53–0.70]
Submucosal infiltrate	0.80 [0.75–0.84]	0.85 [0.81–0.88]
Abscess	0.69 [0.58–0.77]	0.69 [0.60–0.79]
Ulceration	0.48 [0.35–0.59]	0.56 [0.44–0.68]
Total score	0.90 [0.87–0.92]	0.92 [0.89–0.94]

disease was assessed by all observers using a 10-cm visual analogue scale [VAS], in which no disease activity was scored as 0 and the most severe activity as 10.

2.5 Statistical methods

The intra-rater reliability was defined as the correlation between two measurements on the same slide or video by the same observer, while the inter-rater reliability was defined as the correlation between the scoring on the same slide or movie by the two different observers. The intraclass correlation coefficients (ICCs) for both inter- and intra-rater agreement, based on all available slides and videos, were estimated using a two-way random effects model, with interaction between slides or videos and raters using the maximum likelihood method. The ICCs were evaluated using the Landis and Koch criteria, whereby <0.00, 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00 indicate a 'poor', 'slight', 'fair', 'moderate', 'substantial' and 'almost perfect' agreement, respectively.[18]

For the model development process, control animals were excluded and only the first set of observations from one of the observers was used. Exploratory bivariate analyses between the VAS and each of the items selected based on reliability were performed first to guide the coding of each item. Specifically, we

pre-specified that item variables would be coded as continuous if a linear relationship was demonstrated between the item score and the VAS. If a linear relationship was not evident, the bivariate relationships were used to collapse item levels. A full model was then obtained using all items, followed by a step-down model-building approach with $p = 0.05$ used as the criterion for item selection. Residuals from the final model were subjected to statistical diagnostics examination. The stability of the final model was assessed and calibrated using the bootstrap method with 2000 replicates. For ease of calculation, we standardized the regression coefficients by dividing the smallest coefficient and rounding to integers. The results of the first set of observations for the observer not used in the model development process were used for purposes of model validation.

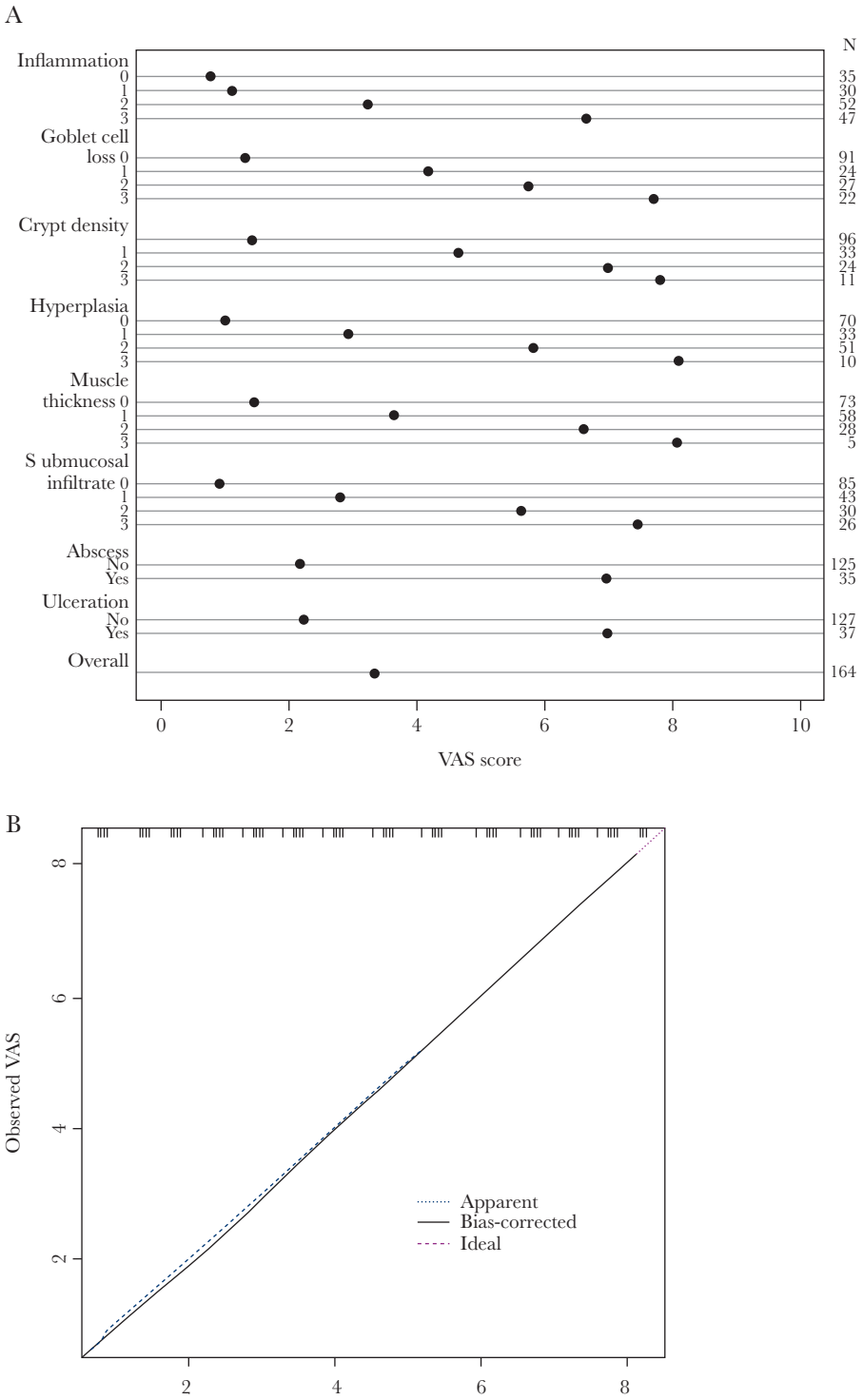

Figure 3. Analysis of the histology items. [A] The univariable summaries of the VAS scores as stratified by the levels of the items in the histology score. [B] Calibration plot of the actual versus the predicted VAS score using the final model with four variables [goblet cell loss, crypt density, hyperplasia, submucosal inflammation]. The ideal prediction is shown by the 45° line, and the model performance [apparent] is shown by the dashed line.

Spearman rank correlations were used to assess convergent validity between the newly developed endoscopic index, the newly developed histologic index, and colon density, using the first score for the endoscopy and histology for each of the observers. External validation was assessed using Spearman rank correlations for the endoscopy VAS with the Mouse Colitis Endoscopy Index [MCEI], and for the histology VAS with the Mouse Colitis Histology Index [MCHI]. Correlations exceeding a threshold of 0.7 were considered acceptable. The ability of the indices to distinguish between controls, untreated and treated mice was assessed using analysis of variance, using a Tukey–Kramer adjustment for pairwise comparisons. Statistical analyses were performed using SAS V 9.4 [SAS Institute, Cary, USA].

3. Results

3.1 Study population

In total we collected 224 histology slides and 201 endoscopy videos from 227 experimental animals, from three independent T-cell transfer experiments and three independent IL10 KO experiments. The number of animals under each experimental condition is given in Table 3.

3.2 Reliability of the histology scoring

The inter-rater and intra-rater reliability coefficients for the eight histologic items are given in Table 4, and within each of the evaluated models in Supplementary Table 1. The inter-rater and intra-rater reliability coefficients and their 95% confidence intervals for the total score were 0.90 [0.87–0.92] and 0.92 [0.89–0.94], respectively, indicating 'almost perfect' agreement. Twenty-seven of the 224 [12.1%] histological slides were reported as being of poor quality. When poor-quality slides were excluded, reliability coefficients were higher for most items [Supplementary Table 2], notably for 'ulceration' and 'abscess', which one might expect to be the most difficult to score reliably on poor-quality slides. When healthy control animals were excluded from the calculations, the reliability coefficients were similar [inter- and intra-rater reliability of 0.88 [0.85–0.91] and 0.90 [0.87–0.93], respectively, Supplementary Table 3].

Table 5. Regression model for the Mouse Colitis Histology Index [MCHI].

	Coefficient [se]	*p*-value
Goblet cell loss	0.371 [0.173]	0.034
0 = none		
1 = <10%		
2 = 10–50%		
3 = >50%		
Crypt density	0.805 [0.214]	<0.001
0 = normal		
1 = decrease of <10%		
2 = decrease of ≥10%		
Hyperplasia	0.584 [0.160]	<0.001
0 = none		
1 = slightly increased crypt length		
2 = 2–3 times increased crypt length		
3 = >3 times increased crypt length		
Submucosal infiltrate	0.996 [0.131]	<0.001
0 = none		
1 = individual cells		
2 = infiltrate[s]		
3 = large infiltrate[s]		

3.3 Histology index development and assessment

The distribution of VAS scores for histology showed that a large percentage was scored as non-diseased [VAS score ≤1, Supplementary Figure S1A], which hampers model building. The VAS scores followed a better distribution when healthy control animals were excluded [Supplementary Figure S1B]. Hence, we excluded the healthy control animals for the model-development process. The bivariate relationships between each of the histological items and the VAS score [Figure 3A] suggest a linear relationship for 'goblet cell loss', 'hyperplasia', 'muscle thickness' and 'submucosal infiltrate'. Thus, the scores for these items were treated as continuous variables in the model-building process. For 'inflammation' there was no clear difference in the VAS for scores 0 and 1, and therefore, these were collapsed into a score of 0, and subsequently 2 was recoded to 1 and 3 to 2. For 'crypt loss' the scores of 2 and 3 were collapsed into a score of 2. Following the step-down procedure with a bootstrap of 2000 resamples, the initial model comprising of all eight items was reduced to a final model with 'goblet cell loss', 'crypt density', 'hyperplasia' and 'submucosal infiltrate' as the items that best predicted the VAS [Table 5], with an R^2 value of 0.89 [R^2 = 0.88 when corrected for optimism]. The calibration plot [Figure 3B] shows that the final model has reasonable external validity. This new index, designated the MCHI, can be calculated as:

MCHI =
1 × Goblet cell loss [four categories]
+ 2 × Crypt density [three categories]
+ 2 × Hyperplasia [four categories]
+ 3 × Submucosal infiltrate [four categories].

The total score of the MCHI ranges from 0 [no disease] to 22 [severe disease]. The inter- and intra-rater reliability coefficients [95% CI] were 0.88 [0.85–0.91] and 0.91 [0.89–0.93], respectively, indicating 'almost perfect' reliability. When the MCHI was applied to the readings from the second observer, the correlation between the MCHI and the VAS was 0.82 [0.78–0.86]. The correlation between the MCHI and colon density was 0.81 [0.75–0.85] for both observers. The MCHI was able to distinguish between the experimental

Table 6. MCHI between-group comparisons.

MCHI [se]	T-cell transfer				IL10 KO			
	Control [*n* = 30]	Diseased [*n* = 46]	Treated[a] [*n* = 23]	[b]*p* =	Control [*n* = 30]	Diseased [*n* = 35]	Treated [*n* = 24]	**p* =
Observer 1	0.17 [0.12]	13.11 [1.12]	1.15 [0.42]	<0.001	0.07 [0.07]	8.60 [1.12]	0.79 [0.34]	<0.001
Observer 2	0.27 [0.12]	13.65 [1.11]	1.22 [0.28]	<0.001	1.23 [0.38]	10.06 [1.13]	2.79 [0.62]	<0.001

[a]Only the effective dose of 100 µg anti-TNFα was included in these comparisons. [b]The *p*-values presented here show the Tukey–Kramer adjusted comparison between the diseased and treated groups. The diseased animals also differed from the control animals [$p < 0.001$ for all], whereas the difference between control and treated animals was non-significant.

Table 7. Reliability coefficients [95% CI] of the endoscopic item scores.

	ICC [95% CI]	
	Inter-rater	Intra-rater
Thickening	0.71 [0.64–0.76]	0.76 [0.71–0.81]
Vasculature	0.58 [0.50–0.66]	0.73 [0.67–0.79]
Fibrin	0.61 [0.54–0.68]	0.74 [0.68–0.79]
Granularity	0.69 [0.63–0.74]	0.77 [0.71–0.81]
Stool[a]	0.69 [0.61–0.76]	0.76 [0.68–0.82]
Total score	0.80 [0.76–0.84]	0.86 [0.83–0.89]

[a]Not available for all subjects, therefore excluded from the total score.

groups within the models, with the pairwise differences between the treated and untreated animals being statistically significant [$p < 0.001$, Table 6].

3.4 Reliability of the endoscopy scoring

The inter-rater and intra-rater ICCs for the five endoscopic items are given in Table 7, and within each of the evaluated models in Supplementary Table 4. Because no feces were visible, and the feces component could not be scored in 43 of 201 [21.4%] of the endoscopy videos, this item was not included in the total endoscopy score. The inter-rater and intra-rater ICCs [95% CI] for the total endoscopy score were 0.80 [0.76–0.84] and 0.86 [0.83–0.89], respectively, representing 'almost perfect' agreement. A small number [8 of 201,

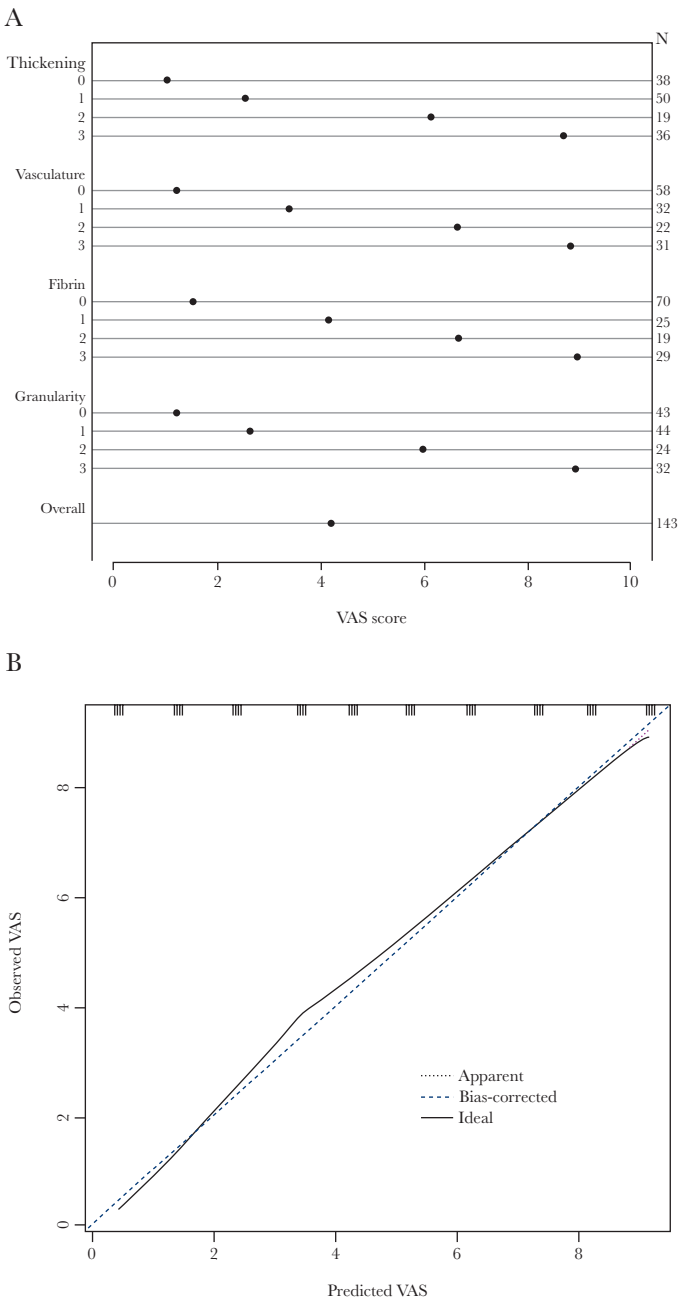

Figure 4. Analysis of the endoscopy items. [A] The univariable summaries of the VAS scores as stratified by the levels of the items in the endoscopy score. [B] Calibration plot of the actual versus the predicted VAS score using the final model with three variables [thickness, vasculature and granularity]. The ideal prediction is shown by the 45° line, and the model performance [apparent] is shown by the dashed line.

4%] of the endoscopy videos were of poor quality. Excluding these videos had little effect on the reliability coefficients [Supplementary Table 5]. When healthy control animals were excluded from the calculations, the reliability coefficients were similar: 0.81 [0.76–0.84] and 0.87 [0.83–0.90] for inter- and intra-rater reliability, respectively [Supplementary Table 6].

3.5 Endoscopy index development and assessment

As was the case with histology, the endoscopy VAS scores showed a better distribution when healthy control animals were excluded [Supplementary Figure 2], and hence, healthy control animals were excluded from the model-development process. The bivariate associations between the four endoscopic item scores and the VAS score exhibited a linear relationship [Figure 4A], and therefore were all treated as continuous in the modeling process. Application of the step-down procedure, with a bootstrap of 2000 resamples, yielded a final model with a combination of 'thickness', 'vasculature' and 'granularity' [Table 8], with an R^2 of 0.87 and good external validity [Figure 4B]. This new endoscopy score, designated the MCEI ranges from 0 to 9 and is calculated as:

MCEI =
1 × Thickening [four categories]
+ 1 × Vasculature [four categories]
+ 1 × Granularity [four categories].

The inter-rater and intra-rater reliability coefficients [95% CI] of 0.80 [0.76–0.83] and 0.85 [0.81–0.88], respectively, indicate 'almost perfect' reliability. When the MCEI was applied to the readings from the second observer, the correlation between the MCEI and the VAS was 0.78 [0.72–0.83]. The correlation between the MCEI and colon density was 0.77 [0.71–0.83] for the first observer and 0.73 [0.66–0.79] for the second observer. The MCEI was able to distinguish between the experimental groups, with pairwise differences between the treated and untreated animals being statistically significant [$p < 0.001$, Table 9]

Table 8. Regression model for the Mouse Colitis Endoscopic Index [MCEI].

	Coefficient [se]	*p*-value
Thickening	0.954 [0.222]	<0.001
0 = transparent		
1 = moderate		
2 = marked		
3 = non-transparent		
Vasculature	0.758 [0.263]	0.005
0 = normal		
1 = moderate		
2 = marked		
3 = absent/bleeding		
Granularity	0.888 [0.210]	<0.001
0 = none		
1 = moderate		
2 = marked		
3 = extreme		

3.6 Correlation of histology and endoscopy

The Spearman rank correlation between the MCHI and MCEI was 0.74 [0.67–0.80] for the first observer and 0.71 [0.63–0.77] for the second.

3.7 External validation for the MCHI and MCEI

To confirm that the newly developed indices were valid in the hands of other researchers, both the histological slides and the endoscopy videos were evaluated by external observers. This external evaluation showed a correlation of the MCHI with the histology VAS of 0.86 [0.82–0.89], providing external validation for our proposed histology index. The definitions of items and scoring criteria for the MCHI were finalized for usage in future studies and are presented in Table 10. For the endoscopy videos, the external evaluation showed a correlation of 0.87 [0.83–0.90] of the MCEI with the endoscopy VAS, providing external validation for our proposed endoscopy index, for which the definitions of included items and scoring criteria were summed and presented in Table 11.

4. Discussion

Endoscopy and histopathology assessments in mouse models for IBD are essential for both fundamental and translational research. Several research groups have developed ordinal classification indices, which to some extent measure overlapping features and generally have face validity [appear to be valid]. However, formal assessment of the operating properties of these indices had not been performed. In the current study, we evaluated histopathology and endoscopy scoring systems for reliability, validity and ability to discriminate differences between groups known to vary in severity in mouse models for IBD, and developed simplified prediction models from these systems. We included two widely used mouse models for chronic intestinal inflammation: the T-cell transfer and the IL10 KO models. Items within the scoring systems were chosen based on the existing indices that are currently used for these mouse models.

Usually an inter-rater ICC of >0.4 is used as criterion for including an item in the development of a new index.[19] We showed that all histology items had a higher reliability than 0.4 [the lowest inter-rater ICC was 0.48, for 'ulceration']. Our observers agreed that this component was the most difficult item to score, especially on poorer quality slides. Indeed, removing the poor-quality slides from the dataset increased the inter-rater ICC of 'ulceration' to 0.53. In addition to

Table 9. MCEI between-group comparisons.

MCEI [se]	T-cell transfer				IL10 KO			
	Control [$n = 30$]	Diseased [$n = 42$]	Treated[a] [$n = 21$]	[b]p =	Control [$n = 28$]	Diseased [$n = 27$]	Treated [$n = 21$]	*p =
Observer 1	0.63 [0.16]	5.43 [0.31]	1.33 [0.31]	<0.001	0.86 [0.20]	4.26 [0.57]	1.19 [0.25]	<0.001
Observer 2	1.10 [0.17]	4.79 [0.39]	1.43 [0.26]	<0.001	1.57 [0.26]	4.37 [0.41]	1.48 [0.21]	<0.001

[a]Only the effective dose of 100 µg anti-TNFα was included in these comparisons. [b]The *p*-values presented here show the Tukey–Kramer adjusted comparison between the diseased and treated groups. The diseased animals also differed from the control animals [$p < 0.001$ for all], whereas the difference between control and treated animals was non-significant.

Table 10. Mouse colitis histology index [MCHI].

Item	Score		Factor
Goblet cell loss	0	none	1
	1	<10%	
	2	10–50%	
	3	>50%	
Crypt density	0	normal	2
	1	<10% decrease in crypt density	
	2	≥10% decrease in crypt density	
Hyperplasia	0	none	2
	1	slightly increased crypt length	
	2	2 to 3 times increase in crypt length	
	3	>3 times increase in crypt length	
Submucosal infiltrate	0	none	3
	1	individual infiltrating cells	
	2	infiltrate[s]	
	3	large infiltrate[s	

Table 11. Mouse colitis endoscopy index [MCHI].

Item	Score		Factor
Thickening	0	transparent	1
	1	moderate	
	2	marked	
	3	non-transparent	
Vasculature	0	normal	1
	1	moderate	
	2	marked	
	3	absent/bleeding	
Granularity	0	none	1
	1	moderate	
	2	marked	
	3	extreme	

'ulceration', the model-building process excluded 'muscle thickness' and 'abscess', items for which the inter-rater ICCs were low.

To enable a better distribution of disease severity scores, which enables better model building, the healthy control animals were excluded from the process of developing a new histological index. This process was started by investigating the bivariate relationships between the VAS score and the individual items. 'Inflammation' and 'crypt loss' exhibited non-linear relationships with the VAS and were re-categorized for the model-development process. Four items ['goblet cell loss', 'crypt density', 'hyperplasia' and 'submucosal infiltrate'] were selected through the modeling process. Weighting of the items was based on the model regression coefficients, and the weighted item scores were summed to generate the MCHI. The MCHI was shown to be able to discriminate between control animals, untreated animals, and animals treated with a treatment of known efficacy. The new histological index has the potential to reduce the sample size required for detecting treatment effects, and as the readers do not have to score all initial features, to reduce the time required to read slides. The reliability coefficients of the histological components were considerably lower for the IL10 KO model than for the T-cell transfer model. As there was a lower number of slides available from the IL10 KO model for the model-development process, and the IL10 KO mice exhibited less severe disease [Table 6], the MCHI may be a more efficient instrument in the T-cell transfer model.

For the five items from the endoscopy scoring system introduced by Becker et al.,[14] the inter-observer reliability was also assessed as 'moderate' to 'substantial'. However, as we tried to remove as much feces as possible before we performed endoscopy, no stool was visible in >20% of the endoscopy videos that were evaluated. Hence, we did not include the stool consistency item in our model development. As was the case with the histology model, control animals were excluded from the model-development process. Consistent with the non-reasoned decision by Norwaski et al., to exclude the fibrin component from their total endoscopy score,[16] 'fibrin' was eliminated during the model-building process. The final model, the MCEI, is comprised of 'vasculature', 'mucosal thickness' and 'granularity', all equally weighted. Correspondingly, 'fibrin' had the weakest correlation with the VAS [not shown].

This study, assessing both endoscopic and histologic semi-quantitative scoring systems in animal IBD models, had several methodological strengths. We used a large number of histological slides and endoscopy videos from two different IBD models. The slides and videos were generated in a standardized manner, using mice from several different experiments that included healthy controls, diseased untreated groups, and therapeutic treatment groups. The inter- and intra-rater reliability coefficients found in this study were generally higher than those in human studies,[19,20] which could be due to the differences between observers, amount of tissue, pathophysiology, and/or the chosen items. The validity of the new indices was tested in various ways. Both the MCHI and MCEI were shown to discriminate between healthy control and [untreated] diseased animals, and therefore both indices possess construct validity [i.e. measure what they are supposed to measure]. Both the MCHI and MCEI correlated with the colon density and with each other, indicating that both indices possess convergent validity [i.e. correlate with other disease parameters], as expected. Moreover, the models for both the MCHI and MCEI were developed based on the results of one observer, and were internally validated by the results of the second observer. Finally, both the MCHI and MCEI were externally validated. Both external observers, one for the endoscopy and one for the histology, showed there was a very strong correlation of the index with the respective VAS, providing external validation for the MCHI and MCEI. In this study, we demonstrated that all histology and endoscopy item scores were at least moderately reliable, suggesting that other indices in the literature based on these items, are also likely to be reliable. However, the operating characteristics of the simple sums of these items, which have commonly been reported as used in the literature, have never been evaluated. The MCHI and MCEI presented here have been configured to be reliable, valid and responsive to therapeutic interventions. The use of these instruments has the potential to enable more efficient pre-clinical studies, specifically in the T-cell transfer and IL10 KO models, enabling the development of new and better therapeutic options for patients with IBD.

Funding

This work was supported by an unrestricted research grant from the Janssen Prevention Center of Janssen Vaccines & Prevention BV, Leiden, the Netherlands[part of the Janssen Pharmaceutical Companies of Johnson & Johnson].

Conflict of Interest

MK and ABvtW are employed by Janssen Vaccines and Prevention B.V. LWS, BGF and GRAMD'H are employed by Robarts Clinical Trials Inc. GRAMD'H has served as advisor for Abbvie, Ablynx, Amakem, Amgen, AM Pharma, Avaxia, Biogen, Bristol Meiers Squibb, Boerhinger Ingelheim, Celgene/Receptos, Celltrion, Cosmo, Covidien/Medtronics, Ferring, DrFALK Pharma, Eli Lilly, Engene, Galapagos, Genentech/Roche, Gilead, Glaxo Smith Kline, Immunic, Johnson and Johnson, Lycera, Medimetrics,

Millenium/Takeda, Mitsubishi Pharma, Merck Sharp Dome, Mundipharma, Nextbiotics, Novonordisk, Otsuka, Pfizer/Hospira, Prometheus laboratories/ Nestle, Protagonist, Robarts Clinical Trials, Salix, Samsung Bioepis, Sandoz, Setpoint, Shire, Teva, Tigenix, Tillotts, Topivert, Versant and Vifor. He has received speaker fees from Abbvie, Biogen, Ferring, Johnson and Johnson, Merck Sharp Dome, Mundipharma, Norgine, Pfizer, Samsung Bioepis, Shire, Millenium/Takeda, Tillotts and Vifor, and has received research grants from Abbvie, Johnson and Johnson, MSD, Medtronics, DrFALK Pharma, Glaxo Smith Kline, Pfizer, Prometheus, Robarts Clinical Trials and Takeda. BGL was an employee of Robarts Clinical Trials Inc, including having been a consultant for Roche, Gilead and Tillots Pharma and having received speaking fees from Prometheus. GRvdB is an employee of GlaxoSmithKline. He has previously received consulting fees from AbbVie and lecture fees from AbbVie, Merck Sharp & Dohme, and Ferring Pharmaceuticals. He has previously received research grants from AbbVie, Crucell, and Ferring Pharmaceuticals.

Acknowledgments

We thank Dr Dave Shealy [Janssen Research & Development] for the therapeutic antibodies and Anouk Gloudemans [Janssen Vaccines and Prevention B.V.] for helpful discussions.

Author Contributions

PJK, MEW, BGF, BGL and GRvdB were involved in the conception and design of the study. PJK, MEW, SD, JFB, AAtV, LWS, RA and MV were involved in acquisition and analysis of data. PJK, MEW, GRAMD'H, ABvtW, MK, LWS, BGL, BGF and GRvdB were involved in interpretation of data and revising the manuscript critically for important intellectual content. PJK and GRvdB drafted the manuscript. All authors approved the final version of the manuscript.

Supplementary Data

Supplementary data are available at *ECCO-JCC* online.

References

1. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. *Vet Pathol* 2013;**50**:1007–15.
2. Erben U, Loddenkemper C, Doerfel K, *et al.* A guide to histomorphological evaluation of intestinal inflammation in mouse models. *Int J Clin Exp Pathol* 2014;**7**:4557–76.
3. Read S, Powrie F. Unit 15.13 Induction of inflammatory bowel disease in immunodeficient mice by depletion of regulatory T cells. In: Coligan JE, National Institutes of Health (U.S.), editors. *Current Protocols in Immunology*. New York: Wiley; 1996.
4. Asseman C, Mauze S, Leach MW, Coffman RL, Powrie F. An essential role for interleukin 10 in the function of regulatory T cells that inhibit intestinal inflammation. *J Exp Med* 1999;**190**:995–1004.
5. Izcue A, Hue S, Buonocore S, *et al.* Interleukin-23 restrains regulatory T cell activity to drive T cell–dependent colitis. *Immunity* 2008;**28**:559–70.
6. Ostanin DV, Pavlick KP, Bharwani S, *et al.* T cell–induced inflammation of the small and large intestine in immunodeficient mice. *Am J Physiol Gastrointest Liver Physiol* 2006;**290**:G109–19.
7. Ostanin DV, Bao J, Koboziev I, *et al.* T cell transfer model of chronic colitis: concepts, considerations, and tricks of the trade. *Am J Physiol Gastrointest Liver Physiol* 2009;**296**:G135–46.
8. Bleich A, Mähler M, Most C, *et al.* Refined histopathologic scoring system improves power to detect colitis QTL in mice. *Mamm Genome* 2004;**15**:865–71.
9. Berg DJ, Davidson N, Kühn R, *et al.* Enterocolitis and colon cancer in interleukin-10-deficient mice are associated with aberrant cytokine production and $CD4^+$ TH1-like responses. *J Clin Invest* 1996;**98**:1010–20.
10. Kullberg MC, Jankovic D, Gorelick PL, *et al.* Bacteria-triggered $CD4^+$ T regulatory cells suppress *Helicobacter hepaticus*–induced colitis. *J Exp Med* 2002;**196**:505–15.
11. Kullberg MC, Jankovic D, Feng CG, *et al.* IL-23 plays a key role in *Helicobacter hepaticus*–induced T cell–dependent colitis. *J Exp Med* 2006;**203**:2485–94.
12. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985;**38**:27–36.
13. Huang EH, Carter JJ, Whelan RL, *et al.* Colonoscopy in mice. *Surg Endosc* 2002;**16**:22–4.
14. Becker C, Fantini MC, Wirtz S, *et al. In vivo* imaging of colitis and colon cancer development in mice using high resolution chromoendoscopy. *Gut* 2005;**54**:950–4.
15. Becker C, Fantini MC, Neurath MF. High resolution colonoscopy in live mice. *Nat Protoc* 2006;**1**:2900–4.
16. Nowarski R, Jackson R, Gagliani N, *et al.* Epithelial IL-18 equilibrium controls barrier function in colitis. *Cell* 2015;**163**:1444–56.
17. Dulai PS, Levesque BG, Feagan BG, D'Haens G, Sandborn WJ. Assessment of mucosal healing in inflammatory bowel disease: review. *Gastrointest Endosc* 2015;**82**:246–55.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
19. Mosli MH, Feagan BG, Zou G, *et al.* Development and validation of a histological index for UC. *Gut* 2017;**66**:50–8.
20. Mosli MH, Feagan BG, Zou G, *et al.* Reproducibility of histological assessments of disease activity in UC. *Gut* 2015;**64**:1765–73.