

Title: Reliability Amongst Central Readers in the Evaluation of Endoscopic Disease Activity in Pouchitis

Authors: Mark A. Samaan, MD^{1,2} Bo Shen, MD³ Mahmoud H. Mosli, MD⁴ Guangyong Zou, PhD^{2,5} William J. Sandborn, MD^{2,6} Lisa M. Shackelton, PhD² Sigrid Nelson, MS² Larry Stitt, MS² Stuart Bloom, MD⁷ Darrell S. Pardi, MD⁸ Paolo Gionchetti, MD, PhD⁹ James Lindsay, MD¹⁰ Simon Travis, MD, PhD¹¹ Ailsa Hart, MD, PhD¹² Mark S. Silverberg, MD, PhD¹³ Brian G. Feagan, MD^{2, 5, 14} Geert R. D'Haens, MD, PhD¹⁵ Vipul Jairath, MD, PhD^{2,5,14}

Author affiliations: ¹Department of Gastroenterology, Guy's & St Thomas' Hospital, London, UK, ²Robarts Clinical Trials Inc., London, Ontario, Canada, ³Center for Inflammatory Bowel Diseases, Digestive Disease Institute, The Cleveland Clinic Foundation, Cleveland, OH, USA, ⁴Department of Medicine, King Abdulaziz University, Saudi Arabia, ⁵Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada, ⁶Division of Gastroenterology, University of California San Diego, La Jolla, CA, USA, ⁷Department of Gastroenterology, University College London Hospital, NHS Trust, London, UK, ⁸Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA, ⁹Department of Medical and Surgical Sciences (DIMEC), University of Bologna-Italy, Bologna, Italy, ¹⁰Bart's Health NHS Trust, The Royal London Hospital, London, UK, ¹¹Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford, UK, ¹²St Mark's Hospital Inflammatory Bowel Disease Unit, Imperial College, London, UK, ¹³Division of Gastroenterology, Mount Sinai Hospital Inflammatory Bowel Disease Centre, Toronto, Ontario, Canada, ¹⁴Department of Medicine, University of Western Ontario, London, Ontario, Canada, ¹⁵Department of Gastroenterology, Academic Medical Center, Amsterdam, Netherlands

Correspondence to: Vipul Jairath, MBChB DPhil, University of Western Ontario, London, Ontario, Canada, OX3 9DU; Tel: 519-663-3655; Fax: 519-663-3658; Email: vjairath@uwo.ca

Keywords: Pouchitis, clinical trials, reliability

Author contributions: MAS, BS, MHM, GYZ, WJS, LS, BGF, GRDH, VJ conception and design; MAS, BS, GYZ, LMS, SN, LS, SB, DSP, PG, JL, ST, AH, MS, BGF, VJ analysis and interpretation of the data; GYZ, LS statistical analysis; MAS, GYZ, LS, LMS, BGF, VJ drafting of the article; all authors provided critical revision of the article for important intellectual content; all authors provided final approval of the article.

Abbreviations: IPAA, ileal pouch-anal anastomosis; UC, ulcerative colitis; PDAI, Pouchitis disease activity index; endoscopic indices; GELS, global evaluation of lesion severity; ICC, intraclass correlation coefficient; CI, confidence interval

Word count: 3645

ABSTRACT

BACKGROUND & AIMS: Pouchitis is a common complication following proctocolectomy with ileal pouch anal anastomosis for ulcerative colitis. Evaluation of pouchitis disease activity and response to treatment requires use of validated indices. We assessed the reliability of items evaluating endoscopic pouchitis disease activity.

METHODS: Twelve panelists used a modified RAND appropriateness methodology to rate the appropriateness of items evaluating endoscopic pouchitis disease activity derived from a systematic review, and also identified additional potential endoscopic items based on expert opinion. Four central readers then evaluated 50 pouchoscopy videos in triplicate, in random order. Intra- and inter-rater reliability for each item was assessed by calculating and comparing intra-class correlation coefficients (ICCs). A Delphi process identified common sources of disagreement among the readers.

RESULTS: Ten existing endoscopic items were identified from the systematic review and an additional 7 exploratory items from the panelists. Intra-class correlation coefficients (95% confidence interval [CI]) for inter-rater reliability were highest for the existing item of pouch ulceration (0.72, 0.60-0.82) and for the exploratory item of ulcerated surface in the pouch body (0.67, 0.53-0.75). Inter-rater reliability for all other existing and exploratory items was “moderate” (ICC<0.60). The item ulcerated surface in the pouch body demonstrated the best correlation with a global evaluation of lesion severity ($r=0.80$, 95% CI 0.73-0.85).

CONCLUSION: Substantial reliability was observed only for the endoscopic items of ulceration and ulcerated surface in the pouch body. Future studies should assess responsiveness to treatment in the next stage towards development of an endoscopic pouchitis disease activity index.

INTRODUCTION

Surgical treatment is required in up to 30% of patients with ulcerative colitis (UC) after a decade of disease,¹⁻³ either as a consequence of medically refractory disease or development of dysplasia. In this situation restorative proctocolectomy with ileal pouch-anal anastomosis (IPAA) is usually the surgery of choice. However, pouchitis may occur within 10 years in up to 50% of patients, and is associated with impaired health-related quality of life due to symptoms of diarrhea, urgency, rectal bleeding or incontinence.⁴⁻⁶ Furthermore, some patients develop chronic pouchitis and either become dependent on antibiotics for symptom relief or have symptoms refractory to conventional therapies. With no approved treatments for this condition, a large unmet medical need exists.

Importantly, several novel therapies are undergoing evaluation in clinical trials, such as the intercellular adhesion molecule-1 anti-sense oligonucleotide, alicaforsen (NCT02525523),⁷ which has been granted orphan designation for this indication by the United States Food and Drug Administration and European Medicines Agency,⁸ and vedolizumab, a monoclonal antibody to the alpha4beta 7 integrin (NCT02790138). However, efficient approaches to evaluation of novel treatments in clinical trials for pouchitis require use of outcome measures with proven validity, reliability and responsiveness.⁹ Although current indices for evaluation of pouchitis typically measure a composite of clinical, endoscopic, and histologic items, none of these, including the most commonly used instrument, the Pouchitis Disease Activity Index (PDAI),¹⁰ were created using robust methodology for outcome measure development.

Pouchoscopy is required both for the diagnosis of pouchitis and to exclude other conditions such as Crohn's disease or structural abnormalities of the pouch. This procedure involves assessment of the rectal cuff, pouch body and pre-pouch ileum. However, standardized and reliable

descriptors of endoscopic disease activity do not currently exist. This study is a first step towards the development of a fully validated endoscopic instrument for the evaluation of pouchitis disease activity. Accordingly, we conducted a systematic review to identify and appraise all evaluative instruments used for the assessment of endoscopic pouchitis disease activity. We then conducted a consensus process using modified RAND appropriateness methodology¹¹ to combine the best available evidence and the clinical experience of experts in the field to rate appropriateness of endoscopic items. The results of blinded central review of pouchoscopy videos were then used to evaluate the reliability of these items to assess endoscopic pouchitis disease activity.

METHODS

Systematic Review of Literature

Search strategy

MEDLINE, EMBASE, PubMed, the Cochrane Library (CENTRAL), and abstracts presented at Digestive Diseases Week and United European Gastroenterology were electronically searched without language restriction from their inception to 2014 to identify endoscopic evaluative instruments used for the assessment of endoscopic pouchitis disease activity. No restriction was placed upon study design, although case studies were excluded. A summary of the specific search strategy used is detailed below and a comprehensive description is included in the Supplementary Material. Each database was searched for (“pouchitis” OR “pouch”) AND (“index” OR “indice” OR “scale” OR “score” OR “grade” OR "Pouchitis Disease Activity Index" OR "PDAI" OR "Objective Pouchitis Score" OR "St. Mark's" OR "Pouchitis Activity Score”).

Study selection

Eligible studies included any study design measuring pouchitis disease activity. Two reviewers (SN and MAS) independently screened citations and abstracts before retrieving full-text publications of all potentially eligible articles. Disagreement was resolved in discussion with a third reviewer (VJ).

The full text of eligible articles was reviewed by pairs of researchers (MAS/SN and VJ/MHM) to extract the following pre-specified variables; index used, disease activity cut-points, whether the index was used for diagnosis, measurement of disease activity or both, as well as study design and number of patients. Additional variables to assess the level of index validation were also collected with regards to the index reliability, validity and responsiveness. Disagreement was discussed amongst individual pairs of researchers and subsequently by all four if agreement was not possible.

Consensus Process

Recruitment of panelist

A panel comprised of 12 international gastroenterologists and a colorectal surgeon, all with a special interest in the care of patients following IPAA, was assembled. The panel included practicing clinicians and inflammatory bowel disease researchers from the United States, Canada, the Netherlands, the United Kingdom and Italy, who were chosen based on their recognized experience.

RAND appropriateness methodology was used to assess the face validity (the extent to which an item is subjectively viewed as addressing the concept it purports to measure) and appropriateness

of items identified in the systematic review to measure endoscopic pouchitis disease activity, as well additional items acquired from endoscopic indices used in the assessment of inflammatory bowel disease activity or considered to be of possible relevance by the experts. RAND appropriateness methodology employs a modified Delphi panel approach to combine the best available evidence with the personal clinical experience of relevant experts.¹¹ The use of a modified Delphi panel to facilitate decision-making is a widely accepted, iterative, evidence-based process.

First-round evaluation of appropriateness

Items measuring endoscopic pouchitis disease activity identified from the systematic review and the additional exploratory items were circulated to panelists in the form of an anonymous, online survey. Definitions acquired from existing, validated scoring systems were supplied for additional clarity where possible.^{12,13} Panelists rated the appropriateness of each item for the measurement of pouchitis disease activity on a scale from 1 to 9 (1 = inappropriate, 9 = highly appropriate).

Panel meeting

Results of the initial survey were distributed to, and discussed with the panelists via a moderated teleconference to identify and examine areas of disagreement on appropriateness of the items and to allow panelists to explain the rationale behind their initial responses. Although this process focussed on detecting consensus amongst panelists, in accordance with RAND appropriateness methodology, no attempt was made to force the panel to consensus.

Minor modifications were made to the questionnaire to improve the clarity of item definitions based on the outcomes of the first panel meeting, and the appropriateness of the modified items to assess endoscopic pouchitis disease activity was then re-evaluated in a second round of panel review.

A final survey was prepared based on discussion of the results of the reliability study, with a focus to identify sources of disagreement among central readers in the interpretation or assessment of endoscopic pouchitis disease activity for items with less than substantial inter-rater reliability. The results of this survey were summarized in the final recommendations of the panelists.

Analysis of panel results

Standardized RAND appropriateness methodology classified each item as *appropriate*, *uncertain* or *inappropriate* for use in assessing endoscopic pouchitis disease activity based on the median panel rating and degree of panel disagreement (1–3 without disagreement = inappropriate; 4–6 OR any median with disagreement = uncertain; 7–9 without disagreement = appropriate).

Disagreement was considered to be present when at least 3 of the 12 panelists rated appropriateness in the 1–3 range and at least 3 others did so in the 7 to 9 range. All items considered appropriate, as well as those that were rated as uncertain, were included for analysis in the subsequent reliability study.

Reliability Study

Study population

A sample of 50 video recordings from patients with UC who had J-pouches and a diagnosis of pouchitis (based upon symptoms, endoscopic, and histological criteria) were prospectively

obtained from the 3 participating centers. With the exception of 3 patients who underwent pouchoscopy for surveillance of dysplasia, the primary indication for pouchoscopy was assessment for active pouchitis (n=47). Indications in these latter patients also included assessment of surgical complication (n=1), dysplasia (n=1) and active cuffitis (n=4). No restrictions were made regarding prior or current treatment. Videos from patients with pouchitis due to other inflammatory etiologies (eg, ischemia or infection), or those from patients with known or suspected Crohn's disease were excluded. All recordings were made using high definition endoscopy equipment. A specialist who did not participate as a central reader in the study reviewed all videos to ensure that a broad spectrum of endoscopic disease activity was available for evaluation, and that videos were of sufficient quality to enable scoring.

Video assessment

Four central endoscopy readers independently reviewed the 50 video recordings, in triplicate, in random order on separate occasions, at least two weeks apart, in the absence of any clinical information. Endoscopic items that were considered appropriate for the assessment of pouchitis disease activity as well as those that were rated as uncertain during the RAND appropriateness exercises were assessed in the pouch body, the pre-pouch ileum, and the rectal cuff as appropriate. Readers also performed an assessment of global endoscopic lesion severity (GELS) on a 10 cm visual analog scale. Video quality was rated as optimal, suboptimal, or unreadable. The intra- and inter-rater reliability for assessment of the endoscopic items and the GELS was assessed by calculating intra-class correlation coefficients (ICCs). Index components with "fair" or "poor" inter-rater reliability based on the criteria of Landis and Koch,¹⁴ whereby ICCs of <0.0, 0.0–0.20, 0.21–0.4, 0.41–0.6, 0.61–0.8, and >0.81 constitute 'poor,' 'slight,' 'fair,' 'moderate,' 'substantial' and 'almost perfect' reliability, respectively, were subsequently discussed during a

consensus meeting of the study readers to identify the most common sources of disagreement for the items. Consensus statements to standardize assessment and minimize variance were generated during this meeting.

Statistical analyses

Descriptive statistics were used to describe the overall rating of disease severity based on the GELS by the central readers. Inferential results, specifically intra- and inter-rater correlations, were presented with point estimates and confidence intervals (CIs). The ICCs for intra- and inter-rater reliability were estimated using a two-way random effects model with interaction between videos and readers. Associated two-sided 95% CIs were obtained using the non-parametric percentile bootstrap method with 2,000 replicates sampled with replacement at the video level to maintain the structure of the data.

Reliability was assessed based on Landis and Koch benchmarks previously described. These empirical benchmarks were originally developed for grading kappa statistics and have become widely adopted for assessment of ICCs. To evaluate construct validity, correlations between the endoscopic items and the GELS were estimated with a mixed model approach using all 12 observations per image made by 4 readers in triplicates. This method was adopted to avoid the potential bias arising from estimates derived from averaging repeated observations. The precision of the estimates was quantified by two-sided 95% CIs obtained using cluster bootstrap methods with 2,000 replicates.¹⁵

The total number of videos required for reliability testing, which did not consider triplicates, thus making it a conservative estimate, was estimated using the method suggested by Zou.¹⁶

Assuming a true ICC of 0.75, rating of 50 videos by a minimum of 4 central readers would yield

an 83% chance of obtaining a one-sided 95% lower boundary that is greater than 0.6, the “substantial” agreement criterion.

Ethical considerations

The use of the pouchoscopy video recordings for the purposes of this study was approved by the University of Western Ontario Research Ethics Board.

RESULTS

Systematic Review

The literature search retrieved a total of 6479 citations, of which a total of 3304 remained after exclusion of duplicates and were screened using predefined eligibility criteria (Supplementary Figure 1). Of these, 3119 articles were ineligible. Eleven articles were obtained through hand-searches and were added, and an additional 60 articles were included based on an updated search through July 2017, resulting in a total of 256 articles for analysis. These articles described a total of 5 disease activity scoring indices (Table 1). The total also included 10 articles that described index derivation and/or studies with some form of index validation (half of which involved correlation of the indices with fecal calprotectin, and the remainder of which correlated indices either with one another, C-reactive protein, or with physician’s assessment). A summary of the endoscopic items derived from those indices, as well as exploratory items thought to be of potential importance that were sent to the panelists for appropriateness rating are shown in Table 2.

Consensus Process

Appropriateness of items

When surveyed on the appropriateness of the endoscopic items identified in existing disease activity indices, the panelists agreed that both *ulcerations* and *erosions* were appropriate for assessment of pouchitis disease activity, and that *edema* was inappropriate (Table 2). The panelists were uncertain regarding the appropriateness of all of the other index items. Of the exploratory items, *stenosis*, *affected surface* and *ulcerated surface* were regarded as appropriate by the panelists, although there was uncertainty regarding the appropriateness of estimating the proportion of *ulcerated surface of the rectal cuff* due to the small area potentially affected. Modification of definitions for items with an uncertain rating were subsequently made based on panelist feedback (see Table 2) and all items (including those deemed as “inappropriate”), as well as 2 additional novel exploratory items (*pre-pouch ulceration* and *inflamed rectal cuff*) were submitted to the panelists for a second round of appropriateness rating. For the items derived from existing indices, the panelists were consistent with their rating of *edema* as inappropriate for assessment of disease activity on the second survey, and concluded that *mucosal flattening* was also inappropriate.

Consistent with the first survey, *erosion* and *ulceration* were again rated as appropriate. *Contact bleeding*, *spontaneous bleeding*, *erythema (dark red)*, and *mucosal hemorrhages* were also considered appropriate. All other items were considered uncertain. For the exploratory items assessed in the second survey, *affected surface* and *ulcerated surface* remained appropriate. The novel item, *inflamed rectal cuff*, was also considered appropriate, whereas all other exploratory

items were regarded as uncertain by the panelists, and there was disagreement on the appropriateness of *pre-pouch ulceration*.

Reliability of the endoscopic items

Intra- and inter-rater reliability for the endoscopic items and the GELS is summarized in Table 3. The ICCs for intra-rater reliability for the items derived from the pouchitis indices were above the benchmark for substantial reliability (>0.61), with the exception of *granularity and nodularity*, for which moderate intra-rater reliability was observed. Intra-rater reliability was greatest for the assessment of *ulcerations* (ICC = 0.78, 95% CI 0.67, 0.87). For the exploratory items, intra-rater reliability ranged from fair (*anastomotic/staple/suture ulceration/erosions*) to nearly perfect (*ulcerated surface in pouch body*), with the majority of the exploratory items exhibiting moderate to substantial intra-rater reliability. Nearly perfect intra-rater reliability was observed for the GELS.

Most of the ICCs for inter-rater reliability of the items derived from existing indices fell within the range for the benchmarks associated with moderate inter-rater reliability, with the exception of *mucopurulent exudate*, for which only fair inter-rater reliability was observed. Similar to intra-rater reliability estimates, substantial inter-rater reliability was observed for the item of *ulcerations*. Intra-class correlation coefficients for the exploratory items were all below the benchmark for moderate inter-rater reliability, with the exception of *ulcerated surface in pouch body* for which substantial inter-rater reliability was also observed (ICC 0.67, 95% CI 0.53, 0.75). Almost perfect reliability was observed for the GELS.

Correlation between the endoscopic items and the GELS

Most of the endoscopic items derived from existing indices showed at least moderate correlation ($r \geq 0.5$; see Table 3) with the GELS, with the exceptions of *erosions* and *mucosal hemorrhages*.

Ulcerations and *erythema* exhibited the strongest correlations with the GELS ($r = 0.68$ and 0.69 , respectively). Of the exploratory items, *affected surface in pouch body* and *ulcerated surface in pouch body* were strongly correlated with the GELS ($r = 0.70$ and 0.80 , respectively), whereas all other items were either weakly or not correlated.

DISCUSSION

Pouchitis is a frequent outcome following IPAA and is associated with impaired quality of life. There is an urgent need for better therapies and at present no approved treatments are available. Drug development may be hampered by the lack of validated instruments, both in terms of identification of appropriate patients for inclusion in clinical trials, as well as for assessment of treatment outcomes. Given the spectrum of clinical presentation, endoscopic evaluation is likely to be the most objective measure of pouchitis disease activity. Despite the availability of several instruments to measure disease activity, there are no standardized descriptors for existing endoscopic items, and none of these instruments have been fully validated. A fundamental aspect of a measurement tool is to demonstrate reliability,¹⁷ defined as the extent to which raters are able to consistently rate disease activity and the degree to which repeated measurements provide similar results.

The PDAI, which consists of a combination of clinical, endoscopic, and histologic assessments (the “pouchitis triad”), is the most commonly used instrument for both diagnosis and measurement of disease activity.¹⁰ Although the endoscopic items in this index (edema, granularity, loss of vascular pattern, friability, mucus exudate, ulceration) were incorporated based on prior research,¹⁸ and have been subsequently and variably included in modifications of the PDAI, the inclusion of these items is based upon potentially relevant but arbitrary findings. Furthermore, no standardized guidelines or descriptors for their identification exist. In this study,

we compiled a comprehensive list of these items, assessed their appropriateness with modified RAND methodology, and tested their reliability. Assessing nine individual endoscopic items from existing indices, we found substantial intra-rater reliability for all items with the exception of granularity/nodularity, for which only moderate reliability was observed. Corresponding inter-rater reliability was lower (moderate) for all items with the notable exception of *ulcerations* in the pouch body, for which substantial reliability was observed. This discordance between intra- and inter-rater reliability is not unexpected since raters are more likely to agree with themselves than one another, however it provides an opportunity to improve item scoring.

Despite their inclusion in existing instruments, uncertainty regarding the appropriateness of several items was evident during the RAND processes. These items included granularity/nodularity, vascular pattern, mucopurulent exudate, and erythema. The presence of pouch granularity was felt to be generally ubiquitous, and thus the relevance of this finding is unknown. Exclusion of granularity/nodularity from future validation and index development efforts was recommended in the post-reliability study Delphi consensus meeting. Furthermore, assessment of vascular pattern was also considered problematic, as loss of vascular pattern is frequently evident in a normal pouch. This observation is likely a consequence of the difficulty associated with visualization of the vascular pattern in the small bowel. The panelists were also divided on the exclusion of the item assessing mucopurulent exudate since this may occur in the absence of inflammation and is highly dependent on washing technique.

Of the items considered appropriate by the panelists (bleeding, erosions, ulceration, mucosal hemorrhages), ulceration was the item with the highest inter-rater reliability. Although the other items were rated as moderately reliable, the panelists recommended removal of mucosal hemorrhages and refinement of the bleeding item to “absent or present” to potentially simplify

and improve its reliability. Furthermore, although intra-rater reliability was substantial for the assessment of erosions, lack of quantitation of the number of erosions was put forward as a potential explanation for the relatively lower inter-rater reliability for this item. Panelists therefore recommended that erosions be categorized as absent or present in the pouch body, and when present, further categorized as <15 or ≥ 15 . However, there is currently no evidence to suggest that the presence of fewer than 15 erosions is associated with fewer symptoms, better outcomes or response to treatment. The reliability of this descriptor should be formally evaluated in future studies. It should also be noted that during the process of refining the items included in our index, the descriptors for the ulceration and erosion items were modified to exclude these findings along the staple/suture line. This recommendation was made by the panel based on the concept that endoscopic changes observed solely at the suture/staple line are generally considered a distinct entity to pouchitis.¹⁹⁻²¹ Mechanical, ischemic and/or local immunological mechanisms have been postulated to contribute to these findings, although no etiology has been confirmed. Nonetheless, endoscopists should avoid misinterpreting isolated changes in this region as evidence of pouchitis. Furthermore, foreign body granulomas have been described, suggesting biopsies should be avoided in this area to avoid diagnostic uncertainty.²²

Several exploratory endoscopic items not currently found in existing indices were also subjected to appropriateness and reliability testing in our study. Intra-rater reliability for the assessment of these items ranged from fair (anastomotic/staple/suture ulceration/erosions) to substantial (affected surface and ulcerated surface in the pouch body), however, only ulcerated surface in the pouch body was associated with substantial inter-rater reliability. The validity of the items assessing ulceration in the pouch body is further supported by the observation that these items were highly correlated with the GELS.

Inter-rater reliability for the remaining exploratory items ranged from slight (anastomotic/staple/suture ulceration/erosions) to moderate (affected surface in pouch body and pre-pouch ulceration) with all other items demonstrating only fair inter-rater reliability. The assessment of pre-pouch ulceration was thought to have been influenced by the location specified in the item descriptor, and a recommendation was made to correct this to 2 cm above the pouch inlet (rather than above the pouch body). This criterion was established to differentiate pouch inlet ulceration from genuine pre-pouch ileitis, two potentially distinct pathological entities. However the validity of this recommendation needs to be established in future studies. Infrequent observation of ulcers in this location may have also contributed to relatively lower reliability for this item compared to the other items assessing ulceration. In the case of the items assessing stenosis and sinus/fistula, infrequent occurrence, and uncertainty of the appropriateness of the items for assessment of disease activity may have resulted in relatively poorer reliability. Moreover, these latter two items were considered post-surgical complications or more frequently associated with Crohn's disease (which was an exclusion criterion for this study) rather than measures of pouchitis disease activity.

In conclusion, we found that the presence and extent of ulceration in the pouch body demonstrated the best inter-rater reliability for the endoscopic assessment of pouchitis disease activity. Items with at least moderate intra- and inter-rater reliability may be further refined to improve both the appropriateness and the reliability of assessment. Future research should also determine the responsiveness of these items, with an ultimate goal of developing a fully validated endoscopic index of pouchitis disease activity for use in clinical practice and in drug development.

Table 1. Summary of endoscopic indices identified in systematic review

Index	Description	Level of validation
St Mark's Score (1977)	Six item numerical grading system, total score range 0 - 6 <ul style="list-style-type: none"> • Loss of vascular pattern • Granularity • Oedema • Mucosal hemorrhages • Contact bleeding • Ulceration 	Not validated
Pouchitis Disease Activity Index (1994)	Six item numerical grading system, total score range 0 - 6 <ul style="list-style-type: none"> • Oedema • Granularity • Friability • Loss of vascular pattern • Mucous exudate • Ulceration 	Partially validated
Heidelberg Pouchitis Activity Score (2001)	Six item numerical grading system, total score range 0 - 12 <ul style="list-style-type: none"> • Oedema (absent = 0, present = 1) • Granularity (absent = 0, present = 1) • Friability (absent = 0, mild = 1, severe = 2) • Erythema (absent = 0, mild = 1, severe = 2) • Flattening of mucosal surface (absent = 0, present = 1) • Ulcerations/erosions (absent = 0, mild = 1, severe = 2) 	Partially validated
Endoscopic Pouch Activity Index (2005)	Six item numerical grading system, total score 0 – 6 <ul style="list-style-type: none"> • Diffuse erythema • Mucus • Friability • Ulcer • Erosion • Granularity 	Not validated
Japanese Diagnostic Criteria for Pouchitis	Stepwise 3-grade scale: <ul style="list-style-type: none"> • Mild <ul style="list-style-type: none"> - Edema - Granularity 	Not validated

(2007)	<ul style="list-style-type: none"> - Loss of vascular pattern - Erythema (mild) • Moderate <ul style="list-style-type: none"> - Erosion - Aphthoid ulcer - Ulceration (solitary or regional) - Friability - Purulent mucous • Severe <ul style="list-style-type: none"> - Ulceration (extensive or multiple) - Erythema (diffuse) - Bleeding (spontaneous) 	
--------	--	--

Table 2. Endoscopic items and definitions from pouchitis disease activity indices identified in systematic review and exploratory endoscopic items: original and modified versions based on consensus

Item	Original definition	Initial panel appropriateness rating	Modified item	Qualifier	Modified definition(s)	Final panel appropriateness rating
Derived from systematic review						
Edema	Marked thickening of the mucosa with blunting of the mucosal folds	Inappropriate	Unmodified	Absent/present	Unmodified	Inappropriate
Granularity and nodularity	Evident granular, nodular variation of the mucosal surface	Uncertain	Unmodified	Absent/present	Unmodified	Uncertain
Friability/contact bleeding	Absent Mild: submucosal hemorrhages or minor bleeding after passage of the endoscope Severe: moderate to severe bleeding before or after the passage of the endoscope	Uncertain (with disagreement on severe definition)	Bleeding	Absent/present	On contact: no bleeding at the site of assessment before, but bleeding seen after contact with the endoscope Spontaneous: bleeding prior to contact with the endoscope	Appropriate
Loss of vascular pattern	Patchy or complete loss of capillary margin	Uncertain	Vascular pattern	Normal/ patchy obliteration/ obliteration	Normal vascular pattern with arborization of capillaries clearly defined or with blurring	Uncertain

					or patchy loss of capillary margins; patchy obliteration of vascular pattern; complete obliteration of vascular pattern	
Mucopurulent exudate	White or yellow deposits on the mucosa unrelated to any bowel preparation	Uncertain	Unmodified	Absent/present	White or yellow deposits on the mucosa unrelated to any bowel preparation and not confined to an ulcer/erosion	Uncertain
Erosions	Tiny (<5 mm) defects in the mucosa, of a white or yellow colour with a flat edge	Appropriate	Unmodified	None/a few/numerous	Tiny (<5 mm) defects in the mucosa, of a white or yellow colour with a flat edge (excluding erosions found along the staple/suture line)	Appropriate
Ulceration	Larger (>5 mm) defects in the mucosa, which are discrete fibrin-covered ulcers or deeper excavated defects in the mucosa with a slightly raised edge	Appropriate	Unmodified	Absent/present	Larger (>5 mm) defects in the mucosa, which are discrete fibrin-covered ulcers or deeper excavated defects in the mucosa with a slightly raised edge (excluding ulcers found along the staple/suture line)	Appropriate
Mucosal flattening	An absence of the bowel folds, villous flattening or shortening	Uncertain	Unmodified	Absent/present	Unmodified	Inappropriate
Erythema	Absent Mild: some increase in	Uncertain	Unmodified	None/mild/marked	Light red: Some increase in colour of mucosa that is likely to	Uncertain (light red) and appropriate (dark red)

	colour of mucosa. Mucosa looks light red Severe: evident increase in colour of mucosa. Mucosa looks dark red				be abnormal Dark red: Red or crimson colour of the mucosa that is similar to blood and is clearly abnormal	
Mucosal hemorrhages	Some spots or streaks of coagulated blood on the surface of the mucosa ahead of the endoscope	Uncertain	Unmodified	Absent/present	Unmodified	Appropriate
Exploratory						
Stenosis	An abnormally narrow lumen of the bowel visualized during endoscopic procedure that can or cannot be passed with an endoscope	Appropriate	Unmodified	None/pouch inlet/pouch body/anastomosis	Unmodified	Uncertain
Sinus/fistula	A cavity or an abnormal communication between two organs or between an organ and the cutaneous surface	Uncertain	Unmodified	Absent/present	Unmodified	Uncertain
Affected surface;	Unaffected	Appropriate	Affected surface	1%-25%/ 26%-	NA	Appropriate

pouch body, pre-pouch ileum, rectal cuff	<50% 50%-75% >75%		assessed in pouch body (any inflammatory changes except ulceration)	50%/51%-75%/>75%		
Ulcerated surface; pouch body, pre-pouch ileum, rectal cuff	None <10% 10%-30% >30%	Appropriate (except uncertain for rectal cuff)	Ulcerated surface assessed in pouch body	1%-25%/ 26%-50%/51%-75%/>75%	NA	
Pre-pouch ulceration	Ulceration (≥ 5 mm) seen in the pre-pouch ileum	Uncertain (with disagreement)	Pre-pouch ulceration (minimum of 2 cm approximately above the pouch)	None/<5 mm erosions/ ≥ 5 mm ulcers	NA	NA
Inflammation of the rectal cuff	Overt bleeding arising from the rectal cuff mucosa	Appropriate	Unmodified	Absent/present	Granularity, nodularity, erosions and/or ulcerations	Appropriate

Table 3. Reliability of endoscopic items for assessment of pouchitis disease activity and correlation with a rating of global lesion severity

	Reliability (95% CI)		Correlation with GELS (r [95% CI])
	Intra-rater	Inter-rater	
GELS	0.89 (0.83, 0.92)	0.80 (0.71, 0.86)	
Items from existing indices			
Granularity and nodularity	0.52 (0.40, 0.63)	0.49 (0.36, 0.61)	0.57 (0.46, 0.66)
Bleeding (presence)	0.62 (0.51, 0.73)	0.57 (0.44, 0.68)	0.60 (0.47, 0.71)
Bleeding (contact vs spontaneous)	0.65 (0.54, 0.74)	0.54 (0.42, 0.64)	0.57 (0.45, 0.67)
Vascular pattern	0.66 (0.56, 0.73)	0.58 (0.44, 0.66)	0.65 (0.58, 0.71)
Mucopurulent exudate	0.62 (0.53, 0.70)	0.39 (0.28, 0.49)	0.50 (0.38, 0.61)
Erosions	0.71 (0.63, 0.78)	0.45 (0.31, 0.56)	0.43 (0.27, 0.56)
Ulcerations	0.78 (0.67, 0.87)	0.72 (0.60, 0.82)	0.68 (0.56, 0.77)
Erythema	0.68 (0.60, 0.75)	0.47 (0.36, 0.57)	0.69 (0.61, 0.75)
Mucosal hemorrhages	0.61 (0.49, 0.71)	0.51 (0.36, 0.63)	0.46 (0.30, 0.60)
Exploratory items			
Stenosis - Any	0.48 (0.26, 0.62)	0.29 (0.12, 0.49)	0.11 (-0.02, 0.26)
Stenosis - Pouch - Inlet	0.55 (0.35, 0.65)	0.29 (0.12, 0.47)	0.11 (-0.04, 0.25)
Sinus/fistula	0.49 (0.25, 0.72)	0.22 (0.13, 0.28)	0.10 (-0.05, 0.23)
Affected surface in pouch body	0.73 (0.64, 0.79)	0.53 (0.39, 0.64)	0.70 (0.62, 0.76)
Ulcerated surface in pouch body	0.84 (0.77, 0.88)	0.67 (0.53, 0.75)	0.80 (0.73, 0.85)
Pre-pouch ulceration	0.62 (0.43, 0.75)	0.45 (0.29, 0.60)	0.27 (0.07, 0.44)
Inflamed rectal cuff	0.61 (0.52, 0.70)	0.37 (0.26, 0.49)	0.37 (0.24, 0.49)
Anastomotic/staple/suture ulceration/erosions	0.34 (0.22, 0.46)	0.17 (0.09, 0.25)	0.22 (0.10, 0.33)

REFERENCES

1. Solberg IC, Lygren I, Jahnsen J, et al. Clinical course during the first 10 years of ulcerative colitis: results from a population-based inception cohort (IBSEN Study). *Scand J Gastroenterol* 2009;44:431-40.
2. Targownik LE, Singh H, Nugent Z, et al. The epidemiology of colectomy in ulcerative colitis: results from a population-based cohort. *Am J Gastroenterol* 2012;107:1228-35.
3. Langholz E, Munkholm P, Davidsen M, et al. Course of ulcerative colitis: analysis of changes in disease activity over years. *Gastroenterology* 1994;107:3-11.
4. Lightner AL, Mathis KL, Dozois EJ, et al. Results at Up to 30 Years After Ileal Pouch-Anal Anastomosis for Chronic Ulcerative Colitis. *Inflamm Bowel Dis* 2017;23:781-790.
5. Tiainen J, Matikainen M. Health-related quality of life after ileal J-pouch-anal anastomosis for ulcerative colitis: long-term results. *Scand J Gastroenterol* 1999;34:601-5.
6. Fazio VW, Kiran RP, Remzi FH et al. Ileal pouch anal anastomosis: analysis of outcome and quality of life in 3707 patients. *Ann. Surg.* 2013; 257: 679-85.
7. Philpott JR, Miner PB, Jr. Antisense inhibition of ICAM-1 expression as therapy provides insight into basic inflammatory pathways through early experiences in IBD. *Expert Opin Biol Ther* 2008;8:1627-32.
8. Jairath V, Khanna R, Feagan BG. Alicaforsen for the treatment of inflammatory bowel disease. *Expert Opin Investig Drugs* 2017;26:991-997.
9. Khanna R, Feagan BG. Through the looking glass: a journey toward optimal endoscopic assessment of disease activity in ulcerative colitis. *Clin Gastroenterol Hepatol* 2013;11:55-6.
10. Sandborn WJ, Tremaine WJ, Batts KP, et al. Pouchitis after ileal pouch-anal anastomosis: a Pouchitis Disease Activity Index. *Mayo Clin Proc* 1994;69:409-15.
11. Brook RH, in K. A. McCormick R, Md.: DHHS/PHS/AHCPR, AHCPR No. 95-0009, 1994. The RAND/UCLA Appropriateness Method. In: Moore SR, Siegel RA, eds. *Clinical Practice Guidelines Development: Methodology Perspectives*. Rockville, Md: DHHS/PHS/AHCPR, 1994.
12. Mary JY, Modigliani R. Development and validation of an endoscopic index of the severity for Crohn's disease: a prospective multicentre study. *Groupe d'Etudes Therapeutiques des Affections Inflammatoires du Tube Digestif (GETAID)*. *Gut* 1989;30:983-9.
13. Travis SP, Schnell D, Krzeski P, et al. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). *Gut* 2012;61:535-42.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
15. Davison AC, Hinkley DV. *Bootstrap methods and their application*. Vol. 1. Cambridge University Press, 1997.
16. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Stat Med* 2012;31:3972-81.

17. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *Journal of Clin Epidemiol* 2006;59:1033-9.
18. Moskowitz RL, Shepherd NA, Nicholls RJ. An assessment of inflammation in the reservoir after restorative proctocolectomy with ileoanal ileal reservoir. *Intl J Colorectal Dis* 1986;1:167-74.
19. Hata K, Ishihara S, Nozawa H, et al. Pouchitis after ileal pouch-anal anastomosis in ulcerative colitis: Diagnosis, management, risk factors, and incidence. *Dig Endosc.* 2017;29:26-34.
20. Shen B, Fazio VW, Remzi FH, et al. Clinical approach to diseases of ileal pouch-anal anastomosis. *Am J Gastroenterol.* 2005;100:2796–2807.
21. McLaughlin SD, Clark SK, Thomas-Gibson S, et al. Guide to endoscopy of the ileo-anal pouch following restorative proctocolectomy with ileal pouch-anal anastomosis; indications, technique, and management of common findings. *Inflamm Bowel Dis* 2009;15:1256-63.
22. Li Y, Wu B, Shen B. Diagnosis and differential diagnosis of Crohn's disease of the ileal pouch. *Curr Gastroenterol Rep* 2012;14:406–13.