Journal of Crohn's and Colitis, 2021, 609–616
doi:10.1093/ecco-jcc/jjaa216
Advance Access publication October 24, 2020
Original Article

Original Article

Expert Consensus on Optimal Acquisition and Development of the International Bowel Ultrasound Segmental Activity Score [IBUS-SAS]: A Reliability and Inter-rater Variability Study on Intestinal Ultrasonography in Crohn's Disease

Kerri L Novak,[a] Kim Nylund,[b,c] Christian Maaser,[d] Frauke Petersen,[e] Torsten Kucharzik,[e] Cathy Lu,[a] Mariangela Allocca,[f,g] Giovanni Maconi,[h] Floris de Voogd,[i] Britt Christensen,[j] Rose Vaughan,[j] Carolina Palmela,[k] Dan Carter,[l] Rune Wilkens[m,n], IBUS Group

[a]Division of Gastroenterology and Hepatology, Department of Medicine, University of Calgary, AB, Canada [b]National Centre for Ultrasound in Gastroenterology, Haukeland University Hospital, Bergen, Norway [c]Institute of Clinical Medicine, University in Bergen, Klinisk institutt 1, Bergen, Norway [d]Outpatient Department of Gastroenterology, Department of Geriatric Medicine, University Teaching Hospital Lueneburg, Lueneburg, Germany [e]Department of Gastroenterology, University Teaching Hospital Lueneburg, Lueneburg, Germany [f]Humanitas Clinical and Research Centre, Rozzano, Italy [g]Humanitas University, Department of Biomedical Sciences, Milan, Italy [h]Gastroenterology Unit, Department of Biomedical and Clinical Sciences. FBF- L.Sacco University Hospital, Milan. Italy [i]Department of Gastroenterology and Hepatology, Amsterdam University Medical Centre, Amsterdam, The Netherlands [j]Department of Gastroenterology, Royal Melbourne Hospital, Parkville, VIC Australia [k]Division of Gastroenterology, Department of General Surgery, Hospital Beatriz Ângelo, Loures, Portugal [l]Department of Gastroenterology, Chaim Sheba Medical Centre, Tel Hashomer, Israel and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel [m]Gastrounit, Division of Medicine, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark [n]Copenhagen Centre for Inflammatory Bowel Disease in Children, Adolescents and Adults, University of Copenhagen, Hvidovre Hospital, Copenhagen, Denmark

Corresponding author: Kerri Novak, MD, MSc, FRCPC, Division of Gastroenterology and Hepatology, University of Calgary, 3280 Hospital Dr. NW, Calgary, AB T2N 4Z6, Canada. Tel.: +1 403-608-3332; email: knovak@ucalgary.ca

Abstract

Background and Aims: Intestinal ultrasound [IUS] is an accurate, patient-centreed monitoring tool that objectively evaluates Crohn's disease [CD] activity. However no current, widely accepted, reproducible activity index exists to facilitate consistent IUS identification of inflammatory activity. The aim of this study is to identify key parameters of CD inflammation on IUS, evaluate their reliability, and develop an IUS index reflecting segmental activity.

Methods: There were three phases: [1] expert consensus Delphi method to derive measures of IUS activity; [2] an initial, multi-expert case acquisition and expert interpretation of 20 blinded cases, to measure inter-rater reliability for individual measures; [3] refinement of case acquisition and interpretation by 12 international experts, with 30 blinded case reads with reliability assessment and development of a segmental activity score.

Results: Delphi consensus: 11 experts representing seven countries identified four key parameters including: [1] bowel wall thickness [BWT]; [2] bowel wall stratification; [3] hyperaemia of the wall [colour Doppler imaging]; and [4] inflammatory mesenteric fat. Blind read: each variable exhibited moderate to substantial reliability. Optimal, standardised image and cineloop acquisition were established. Second blind read and score development: intraclass correlation coefficient [ICC] for BWT was almost perfect at 0.96 [0.94–0.98]. All four parameters correlated with the global disease activity assessment and were included in the final International Bowel Ultrasound Segmental Activity Score with almost perfect ICC (0.97 [0.95–0.99, p <0.001]).

Conclusions: Using expert consensus and standardised approaches, identification of key activity measurements on IUS has been achieved and a segmental activity score has been proposed, demonstrating excellent reliability.

Key Words: Ultrasound; monitoring; activity index; reliability

1. Background/rationale

The use of objective measures of disease activity for monitoring inflammatory bowel disease [IBD] to guide clinical management is now standard of care, given the disconnect between patient symptoms and disease progression.[1] Cross-sectional imaging is non-invasive, and thus increasingly recognised as an important diagnostic and monitoring tool in Crohn's disease [CD].[2] Intestinal ultrasonography [IUS] is of particular interest, given easy repeatability, lack of required preparation, and low cost.[3–5] In addition, patient preference is an important driver.[6,7] Recent ECCO-ESGAR guidelines highlight both magnetic resonance enterography [MRE] and IUS as first-line modalities for small bowel disease assessment in newly diagnosed CD patients, given their accuracy and lack of ionising radiation exposure.[5,8] Although MRE exhibits similar accuracy to IUS in detection of small bowel disease, it may underperform in colonic disease.[6,9] During routine, regular, intermittent follow-up, MRE use is also limited by lengthy acquisition times and somewhat poor patient acceptance, in addition to high costs.[10] Thus, unlike computed tomography [CT] or MRE, IUS can be easily performed as a point of care scan [POCUS] by gastroenterologists to allow timely provision of disease activity assessment to guide therapy and clinical decisions.[11]

Prospective multicentre clinical studies demonstrating utility are paramount to increase acceptance of IUS as a monitoring tool in the management of IBD. Inclusion of IUS depends on the use of standardised, validated, reliable activity measures reflecting disease activity, also demonstrating therapeutic response and healing. These sonographic measures must be repeatable, consistent, and prospectively defined. A recent meta-analysis evaluated published ultrasound activity scores, concluding that most have significant limitations, and none have been adequately validated.[12] Furthermore, a rigorous description of a standardised acquisition approach and a detailed approach to measurement of key parameters are lacking. Adoption of standardised activity scores, with prospective validation, is also important for expanded use in clinical trials and wider clinical adoption. This is an important gap in the extant literature.

Depiction of inflammation of the bowel on ultrasound is complex and, like other modalities, interpretation requires training and expertise. European guidelines [EFSUMB] exist, outlining features of inflammation reflecting a range of potential IUS parameters contributing to inflammatory activity on IUS in CD.[13] This aligns with similarly variable activity parameters included in published studies.[4,14–16] However, to date there are few published data available on expert consensus regarding the use of these parameters or on the inter-rater reliability of these inflammatory activity parameters between readers of IUS.[17] Therefore, the aim of this study is to establish the core parameters contributing to active intestinal inflammation in CD detected with IUS through expert consensus, to evaluate the inter-rater reliability of these measures through blinded expert reads, and to propose a segmental activity score for luminal CD using reliable intestinal ultrasound parameters.

2. Study design

A phased study design was implemented for this publication. This project reflects the contribution of multiple experts performing IUS in nine different countries [Australia, Canada, Denmark, Germany, Israel, Italy, The Netherlands, Norway, and Portugal]. All acquired images were part of routine or planned scans not solely intended for our study, and were de-identified and collected after receiving patient consent. Given the lack of intervention/direct patient interaction for this evaluation, no multicentre institutional ethics approval was indicated.

2.1. Identification of key sonographic parameters

2.1.1. Phase I: modified Delphi consensus on sonographic parameters of inflammation

Using a modified Delphi process, experts ranked parameters considered imperative for the depiction of disease activity. Three steps were undertaken: first, a comprehensive list of contributing parameters for inflammatory activity, based on a previous review combined with all expert's experience and previous publications, was developed [AA, CM, DC, EC, FP, GM. KNo, KNy, RG, RW, TK][4]; second, a blinded rank order list was generated by the 11 experts after collating results from an electronic survey using a five-point Likert scale on the importance of each measure in assessing disease activity; third, after blinded ranking, all parameters were discussed in plenum with arguments for and against inclusion. A final vote [two options for each parameter: include or exclude] enabled consensus regarding the most important contributing inflammatory parameters required on IUS to measure CD activity.

2.2. Inter-rater reliability

2.2.1. Phase II, first blinded de-identified case read

Twenty established CD de-identified cases were collated from expert centres, including still images and videos including colour Doppler imaging [CDI]. Nine readers [AA, CL, CM, DC, FP, GM, KNy, RW, TK] participated in blinded review of cases in PC format embedded into a PowerPoint presentation and rated for activity parameters according to phase I, plus image quality and rater confidence [both on a five-point Likert scale]. Physician global assessment of disease activity and disease severity, based on the available images/videos, were also rated (both on a visual analogue scale [VAS] from 0 to 100). Global disease activity was considered from 0 [normal disease] to 100 [most active disease ever seen], based upon the parameters included in phase 1. All findings were independently entered into a REDCap database. Inter-rater reliability was calculated for each parameter. Case by case review of findings to identify variation based on individual measures was undertaken to clarify disagreement. Standardised still image and cineloop acquisition instructions were collectively developed to improve consistency of case collection, in addition to measurement of activity.

2.2.2. Phase III, final blinded, de-identified case read

Following the establishment of case acquisition and interpretation, a final read of 30 new, consistently acquired [cineloop length, scan plane, CDI box size and settings], de-identified CD cases covering the full spectrum of luminal inflammation from normal to severe activity, including all segments of the bowel, was undertaken. Twelve IUS experts [BC, CP, CL, CM, FdV, FP, KNo, KNy, MA, RV, RW, TK] from eight countries completed the read. Data were available as DICOM cineloops for all central readers, allowing exact caliper measurements as on the US machine itself. Again, all cases were independently [blinded] rated and entered into REDCap without additional history regarding symptoms/additional imaging, endoscopy, or other investigations. Individual activity parameters were reported by all blinded readers, with the same grading as the previous read. In addition, readers were again asked to score the global disease activity on a VAS from 0–100.

2.3. Data collection and analysis

Initial phase 1 data were collected using Mentimeter® online blinded voting. Phase II and III data were consistently collected within REDCap® databases, accessible upon invitation to all experts electronically and password-protected. Missing data were identified, and reminders sent to all contributors. Statistical analysis was performed using Stata/SE 16.1 for Mac [Stata Corp LP, College Station, TX]. Continuous data (bowel wall thickness, global disease activity, and International Bowel Ultrasound [IBUS] score) variability were compared using intraclass correlation coefficient [ICC] based on a mean rating [k = 12], absolute agreement, and a two-way mixed-effects model. Categorical variables were compared using weighted Fleiss' kappa and interpreted based on Landis & Koch benchmarks.[18,19] For development of the new index, items with at least a moderate level of reliability [mean of the two scoring rounds] were selected as candidate items in developing a new index. Multiple regression analysis was undertaken with these variables to create a segmental activity score [SAS] based on the global disease activity assessment. Regression coefficients were used to build the final IBUS segmental activity from 0 [theoretical no disease] to 100 [theoretical worst ever activity/disease] and omitting the β_0 constant/intercept and capped at 100 for final index simplification. The IBUS-SAS score was then calculated for each case, using individual raters' item scores. Reliability of the score expressed as ICC were finally computed. A p-value of <0.05 was considered significant.

Table 1. Complete list of activity parameters derived from expert consensus.

Bowel wall thickness [BWT]	Mucosal ulcers
Colour Doppler imaging signal [CDS]	Length of disease
Inflammatory mesenteric fat [i-fat]	Disease location
Bowel wall stratification [BWS]	Intraperitoneal free fluid
Complications [stenoses, fistulae]	Serosal margin spiculation
Abnormal peristalsis	Mesenteric lymph nodes

3. Results

3.1. Phase I, modified Delphi consensus

The experts identified an exhaustive list of 12 parameters contributing to inflammation [Table 1]. All parameters were ranked. After the first round of voting and subsequent discussion, the parameters were narrowed to include four key parameters: bowel wall thickness [BWT], colour Doppler signal [CDS], inflammatory fat [i-fat], and bowel wall stratification [BWS]. The strongest and most important parameter based on expert consensus and extant literature is BWT, with a threshold of pathology established at >3.0 mm. The same parameters for all segments of the colon and small bowel were adopted. For assessing bowel wall vascularity, a modified Limberg score was adopted [Table 2 and Supplementary data 1, available at *ECCO-JCC* online].

3.2. Phase II, first blinded read

The first blind read of 20 cases was completed by nine IUS experts. Since data were presented in PowerPoint in PC format, BWT could not be measured in a reliable way and was therefore not performed. The inter-rater reliability for the other parameters was moderate to substantial, 0.45–0.62 [Table 3]. Understanding the identified variability contributed to discussion and then agreement on image and cineloop acquisition, outlining a consistent approach to measurement of all parameters including the necessity of proper evaluation of bowel wall thickness [Figure 1 and Table 2; and Supplementary data 2,] available at *ECCO-JCC* online. Ultrasound machine settings were included, to ensure consistency in IBUS CDS [modified Limberg] scoring and it was determined that recordings should be available in DICOM format, allowing reliable distance measuring.

3.3. Phase III, second blinded read

There were 12 IUS experts [Supplementary data 3, available at *ECCO-JCC* online] who completed the second blinded read of 30 established CD cases, using the new acquisition method for the third and final phase of this project. Inter-rater reliability was almost perfect for BWT (95% confidence interval [CI]: ICC = 0.96 [0.94–0.98], p <0.001) and there was moderate agreement for CDS κ = 0.60 [0.48–0.72], p <0.001. Agreement for i-fat detection was also moderate with κ = 0.51 [0.34–0.67], p <0.001, and for BWS was fair with κ = 0.39 [0.24–0.53], p <0.001 [Table 3]. The 'uncertain' category did not occur frequently, with 18% and 7% of cases scoring i-fat and BWS, respectively. There was no significant change in the agreement for any parameter between the first and second phase of this consistency study [Table 3]. Confidence in interpreting IUS images was

Table 2. Core activity parameters, Delphi grading consensus

	Normal	Uncertain	Activity	
BWT	≤3 mm	NA	>3 mm	
i-fat	0 = Absent	1 = Uncertain	2 = Present	
CDS	0 = Absent [none]	1 = Short signals	2 = Long signals inside bowel	3 = Long signals inside & outside bowel
BWS	0 = Normal	1 = Uncertain	2 = Focal [≤ 3 cm]	3 = Extensive [>3 cm]

BWT, bowel wall thickness; i-fat, inflammatory fat; CDS, colour Doppler signal; BWS, bowel wall stratification; NA, not applicable.

Table 3. Expert-derived blinded voting results: inter-rater reliability for IUS parameters during first and second round of voting.

	Coefficient 1st round	Coefficient 2nd round	*p*-value
BWT	NA	0.96 [0.94–0.98]*	NA
CDS	0.62 [0.42–0.82]	0.60 [0.48–0.72]	0.776
i-fat	0.45 [0.27–0.64]	0.51 [0.34–0.67]	0.531
BWS	0.50 [0.29–0.71]	0.39 [0.24–0.53]	0.120
Confidence	0.06 [0.0–0.16]	0.08 [0.0–0.17]	0.534
Quality	0.15 [0.05–0.25]	0.14 [0.04–0.23]	0.776
Activity	0.92 [0.82–0.98]	0.96 [0.94–0.98]*	0.005
Severity	0.97 [0.91–0.99]	0.93 [0.87–0.97]*	0.980

Parameters in bold are included in the final international bowel ultrasound segmental activity score [IBUS-SAS]. All measures are weighted Fleiss' kappa, except demarked with * = intraclass correlation coefficient [ICC] based on a mean-rating [k = 12], absolute agreement, two-way mixed-effects model.[18]

BWT, bowel wall thickness; i-fat, inflammatory fat; CDS, colour Doppler signal; BWS, bowel wall stratification; NA, not applicable.

lower for the second read versus first read on a five-point Likert scale at 3.14 [2.99–3.28] and 3.48 [3.31–3.65], respectively [p <0.004]. The same was true for image quality assessment at 3.08 [2.90–3.28] and 3.51 [3.31–3.72] respectively [p <0.004]. There was close correlation between perceived case quality and confidence rating: r = 0.91 [p <0.0001, Figure 2]. Mean ratings of all 30 cases are shown in Supplementary data 4, available at *ECCO-JCC* online.

3.4. International Bowel Ultrasound Segment Activity Score [IBUS-SAS]

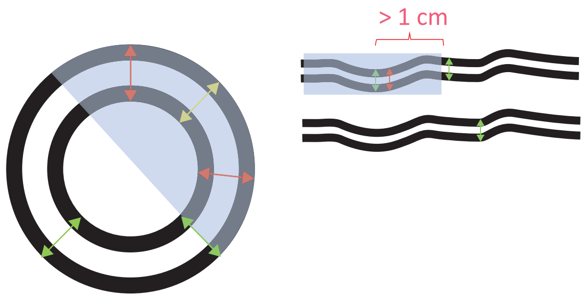
Based on mean ratings from all 12 expert raters, the correlation coefficient between global disease activity defined on a VAS scale and each of the four individual items; BWT, CDS, i-fat, and BWS were computed as 0.73 [0.51–0.87], 0.85 [0.71–93], 0.93 [0.86–97], and 0.87 [0.74–94], respectively [all were significant with a p-value <0.0001, Figure 3]. Based on a multiple linear regression model [Table 4], the following score was developed using ordinal values from Table 2 and BWT in millimetres to reflect global disease activity:

$$\mathrm{IBUS-SAS}\,(0-100) = 4\cdot\mathrm{BWT} + 15\cdot\mathrm{i\text{-}fat} + 7\cdot\mathrm{CDS} + 4\cdot\mathrm{BWS}$$

When the IBUS-SAS was applied to each rating using original scores, ICC for overall IBUS-SAS was almost perfect at 0.97 [0.95–0.99] [p <0.001]. An example of score application is shown in Figure 4.

4. Discussion

This expert consensus on inflammatory activity parameters, combined with a blinded agreement study, is the first centrally read



Yellow arrow represents first measurement. Green arrows are allowed second measurement, red arrows are false second measurements

Figure 1. Measurement of bowel wall thickness. Measures of the bowel wall occur in two orientations: cross-section and longitudinal. The calipers are placed perpendicular to the wall, with two individual measures taken in each orientation, at least 1 cm apart in longitudinal and more than 90° in cross section, in the segment of bowel most affected by disease. The caliper placement occurs from the interface of the mucosa and muscularis mucosae, to the serosa [interface between the serosa and muscularis propria]. All four measures are averaged. Yellow double-headed arrow is the first measurement. Green double-headed arrows are valid second measurements, where red double-headed arrows are invalid caliper placements.

international collaboration of its kind known to date. Activity parameters were selected by 11 experts, based on demonstration of expertly perceived significance in addition to existing evidence supporting association with disease activity, in combination with reliability and interobserver agreement from earlier studies.[4,20–24] BWT is consistently established as the most important predictor of disease activity on ultrasound,[25] and here demonstrates almost perfect inter-rater agreement and correlation with overall assessment of disease activity. CDS, i-fat, and BWS are also important parameters, showing moderate or fair inter-rater agreement and even stronger association with overall assessment of disease activity. In our rigorous attempt to standardise measurement, we endeavor to optimise acquisition and measurement techniques, with an aim to limit uncertainty in the interpretation and grading of the individual parameters.

The importance of BWT as a measure of inflammatory activity for all cross-sectional imaging modalities cannot be understated. When considering well-established indices for MRE, including the Magnetic Resonance Index of Acttivity [MaRIA] score, BWT is a core component.[26] Similarly for CT, BWT and hyperenhancement are the strongest predictors of disease activity.[27] BWT is also the core element in two newly developed simple IUS scores.[25,28] The reliable measurement of BWT is central to consistent interpretation. Again, the reliability of BWT demonstrated here was almost perfect.

The primary focus of this work was to develop and describe an expert consensus approach to measurement techniques and image/cineloop [video] acquisition [see Supplementary data 2], where reliable

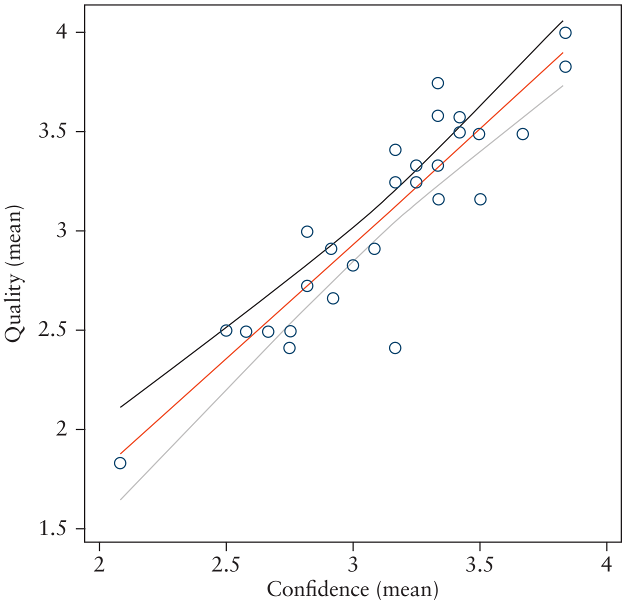


Figure 2. The association between scan quality and rater confidence. Quality and confidence scored on a 5-point Likert scale. Red line is the linear association, grey lines are confidence intervals.

components then allowed for the development of an activity and severity IUS score. This has been poorly demonstrated in the literature to date.[4,12] However, a similar process has been well established by our Rheumatology colleagues. A clear and expert endorsed process of acquisition is essential to training and reproducibility, but this has not yet been developed and published for IUS.[29] For example, Calabrese *et al*. [2018] investigated the inter-rater reliability by comparing IUS interpretation by six independent expert operators evaluating 15 live Crohn's patients, and demonstrated moderate agreement for BWT, BWS, and CDS.[17] Agreement was substantial for lesion location, presence of complications including fistulae, and penetrating complications such as inflammatory masses. However, insufficient agreement was observed for other parameters, such as inflammatory mesenteric adipose tissue [i-fat] changes. Although the investigators evaluated the reliability of key measures of disease activity, little attempt was made to understand the underlying inconsistencies driving scoring differences, which may contribute to their findings of moderate agreement overall. Furthermore, no consensus regarding image/ cineloop acquisition was developed, nor was scoring established before test examination. Demonstrating reliability is paramount for consistent, standardised measurement uptake of ultrasound internationally, to guide therapeutic intervention. In addition, reliable measures are essential for increasing interest in IUS utility in pharma-sponsored therapeutic clinical trials, given its excellent patient acceptance and low cost.[30]

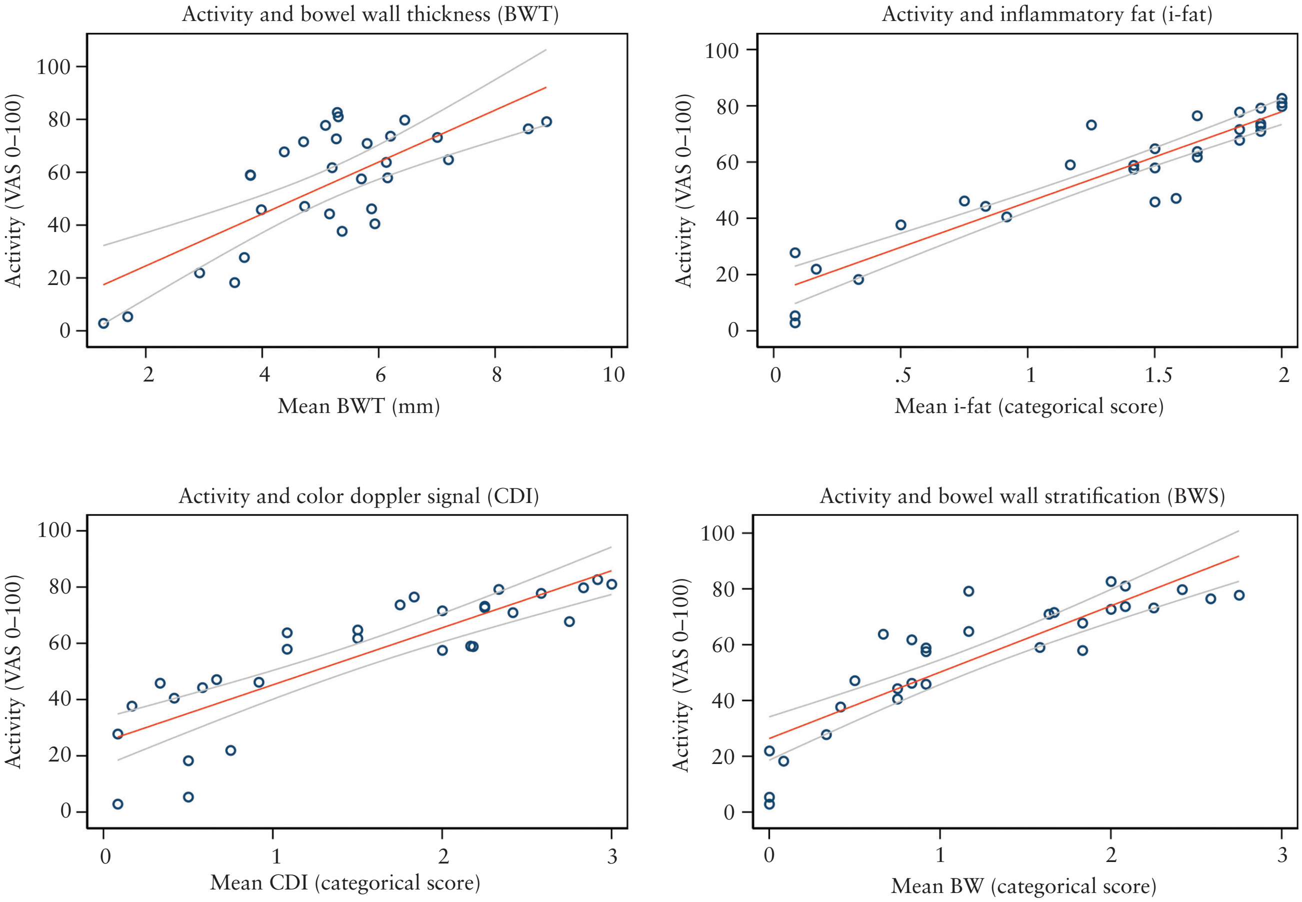


Figure 3. The association between physician global disease activity assessment and individual intestinal ultrasound parameters. Associations between A] mean activity and bowel wall thickness [top left], B] mean activity and inflammatory fat [top right], C] mean activity and colour Doppler imaging [bottom left], and D] mean activity and bowel wall stratification.

Table 4. Multiple linear regression coefficients included in the final activity score.

ACTIVITY		
Parameter	Coefficient	p-value
BWT	4.0 [3.1–4.9]	0.001
i-fat	14.8 [9.8–19.8]	0.001
CDS	6.7 [3.3–10.0]	0.001
BWS	4.1 [0.3–7.9]	0.034

Regression coefficients were calculated based on the global activity and severity score from 0 [theoretical no disease] to 100 [theoretical worst ever activity/disease] and omitting the β_0 constant/intercept. Total model has an F-value of 0.0001 with an adjusted R^2 = 0.99 for all models.

BWT, bowel wall thickness; i-fat, inflammatory fat; CDS, colour Doppler signal; BWS, bowel wall stratification; NA, not applicable.

Shared understanding and agreement regarding interpretation and scoring to improve reliability is equally important for endoscopy. Daperno *et al.* [2017] demonstrated improvement in inter-rater reliability during endoscopic scoring of CD using the CD Endoscopic Index of Severity [CDEIS], after discussion and review of score discrepancy, resulting in substantial improvements in agreement.[31,32] Variability in lesion interpretation on endoscopy is well known.[33] For example, when comparing centrally read versus 'on site' readers for endoscopy, consequential differences in treatment effects were demonstrated by Feagan *et al.* in both placebo and treatment groups.[34] Although revision and refinement of image and cineloop acquisition in this study did not significantly alter the reliability in the final interpretation, standardised acquisition may contribute to improved reliability in less experienced readers and is also important for establishing training standards.

Widespread adoption of IUS has been somewhat limited to date, at least in part due to the common perception that accuracy and thus utility depend on expertise for acceptable performance. Reliance on expertise/skill for optimal performance is no different when comparing IUS with the specialised skills required for the interpretation of MRE and for CTE; in fact, the evidence presented in this study suggests some aspects of IUS measurement may be more reliable.[31,35] For example, Tielbeek *et al.* [2013] investigated inter-rater reliability for MRE activity parameters for CD by two experienced and two less experienced raters.[36] The agreement for most activity parameters ranged from only fair to moderate. Bowel wall thickness, which is perhaps the most used parameter in IUS to demonstrate disease activity, exhibited good inter-rater reliability among all readers for MR [ICC = 0.69]. Experienced radiologists exhibited excellent agreement [ICC = 0.87], yet this was still not as reliable as the findings demonstrated on IUS in this study. Tsai *et al.* [2019] also clearly demonstrated the need for expertise to improve consistent detection of inflammatory parameters on MRE in Crohn's disease among general radiologists.[35] When internationally recognised experts interpret MRE, the agreement is strong, again supporting training and expertise as a component of strong inter-rater reliability for MRE.[9] Similarly, all of the readers in our IUS study were highly experienced, with a median of 6 years of routine clinical use. Not all parameters measured on MR exhibit consistent reliability, regardless of expertise: one of the core elements of the MaRIA score, mucosal ulceration, exhibits less than expected reliability in this real-world assessment. Similarly, the reproducibility of inflammatory fat on MRE demonstrates only fair agreement in Tielbeek *et al.* CT enterography interpretation also requires advanced skill, with some parameters exhibiting greater reliability than others: interobserver agreement reached a kappa of 0.83 for BWT, whereas other parameters like mural hyperenhancement, stratification, fat attenuation, and comb sign were less reliable, with kappas ranging from 0.56 to 0.65.[37] Thus compared with both MR and CT, the reliability of the core parameters [CDS and i-fat] exhibited nearly equivalent reliability and the most important parameter, BWT, was superior.

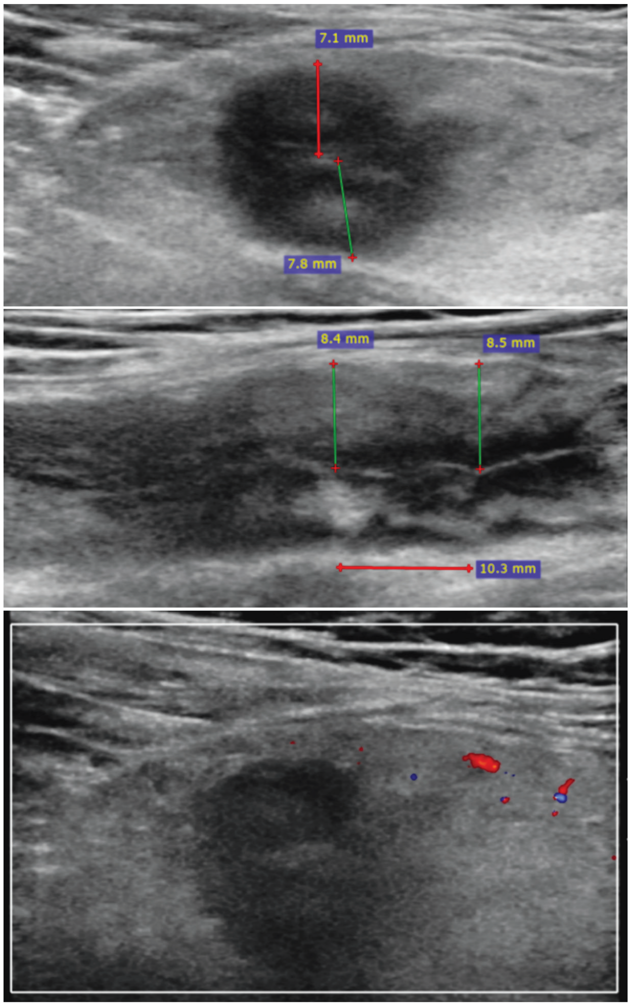


Figure 4. Application of the segmental activity and severity scores. Applying the scores: Bowel wall thickness [BWT] = [7.8 + 7.1 + 8.5 + 8.4] / 4 = 7.95 ≈ 8.0. Blood flow/ colour Doppler signal [CDS] = 0 [no signals]. Inflammatory fat [i-fat] = 2 [certain]. Bowel wall stratification [BWS] = 2 [focal disruption <3 cm]. International Bowel Ultrasound [IBUS] Segmental Activity Score [SAS] = 8 · 4 + 2 · 15 + 0 · 7 + 2 · 4 = 70.

This study has demonstrated that four variables [BWT, BWS, CDS, and i-fat] are required to predict overall disease activity in CD and just two variables [BWT and i-fat] are needed to predict overall disease severity. We have proposed a score that predicts overall disease activity in large and small bowel disease based on a VAS evaluation, although further external validation is required. Use of a physician global assessment or VAS for interpretation of any dependent variable is a common statistical approach in index generation, where no relevant gold standard exists. For example, for clinical symptom measurement in CD [CDAI][38] [overall evaluation of severity of illness] a VAS was implemented in development. Similarly the CDEIS[39] [Crohn's Disease Endoscopic Index of Severity] used a VAS for the global evaluation of lesion severity, and the Lemann score,[40] known to evaluate the overall damage secondary to CD, used a linear VAS. Although our VAS evaluation was performed by the same investigators as those rating the independent variables, it is a different measurement, given the need to include an overarching or 'global' assessment. In comparison with other diagnostic scores, our score may be more comprehensive.[25,28] For example, we have incorporated an appropriate sample size and patient selection, blinded to patient disease characteristics and treatment, and included patients with varying disease activity. BWT was measured as a continuous variable as opposed to categorical, thus making the scores more likely to be responsive. The implementation of central reading allows for objective, standardised grading of images by trained graders. To our knowledge, our stringent methodological design makes our IBUS-SAS the most comprehensive US index currently available for grading CD activity. Even though the reliability of the individual parameters [ranging from fair to excellent] is not perfect, our methods demonstrate high apparent reliability of the final index; yet it does need to undergo external validation against other objective and valid anchors for disease activity. The IBUS-SAS may be used in the future, to predict responsiveness and outcomes following treatment, and index responsiveness should be addressed in the future. When validated against recognised modalities, namely endoscopy and/or MRE, these scores will contribute to standardised measurement in daily clinical practice and will be instrumental in assessing disease activity and response in future clinical trials.

There are a number of limitations to this phased evaluation. First, we did not perform a rigorous systematic review in order to identify IUS parameters evaluating disease activity in CD, but rather relied on expertise and subsequent expert consensus. A recently published systematic review did however reveal only two additional parameters: bowel compressibility and bowel wall echogenicity.[21] Second, the phases of this project occurred over an extended period; thus, some minor variation in participation of experts occurs between blind reads. However, the consistent presence of six core individuals forming the majority readers persisted throughout the project. This investigation was also limited to the evaluation of the inter-rater reliability of parameters for cases read centrally, excluding the evaluation of case reproducibility and intra-rater reliability. Undoubtedly, poor acquisition skills may worsen the reliability between examiners. Attempts were made to mitigate this through generation of clear acquisition instruction. This is a potential limitation of central reading. The aim of this study, however, was to evaluate the reliability of interpretation and grading of images and cineloops followed by score development, not on evaluation of IUS performance nor internal bootstrap validation of the model. There was little to no capacity within this current study to evaluate the consensus score's responsiveness to medical therapies. This will be evaluated in the subsequent phase of score development, in addition to prospective external validation.

In conclusion, IUS is increasingly being adopted as a patient centred, easily repeated objective monitoring tool for CD. Standardised measurement, with consistent acquisition, and interpretation are key to broader application of IUS. The single most important parameter, BWT, exhibits near perfect agreement here, and the other activity parameters demonstrate fair to moderate reliability. The proposed IBUS score associated with global activity also demonstrates excellent reliability. Improved acquisition, measurement, and thus interpretation will facilitate broader inclusion of IUS both in everyday practice and in clinical trials, as we focus on more patient-favoured monitoring tools in CD.

The data underlying this article are available in the article and in its online Supplementary material for the second round of scoring. The data underlying this article will be shared on reasonable request to the corresponding author for the first round of scoring.

Funding

This work was fulfilled without funding support.

Conflict of Interest

KNo reports advisory board fees from AbbVie, Janssen, Pfizer, Ferring, and Takeda; speaker's fees from AbbVie, Janssen, and Pfizer; and research support from AbbVie and Janssen. CM reports advisory and consultation fees from AbbVie, Janssen, Pfizer, Ferring, Takeda, and Roche. CL reports consulting or advisory board fees from AbbVie, Ferring, Janssen, and Takeda. MA received consulting fees from Nikkiso Europe, Mundipharma, Janssen, AbbVie, and Pfizer. FV received speaker's honoraria from Janssen. GM received consulting fees/speaker's honoraria from Alfa Sigma, Janssen, AbbVie, Roche, and Gilead. TK received consulting fees and advisory board fees from Janssen, Takeda, and AbbVie. DC has received speaker's fees and/or research support from Takeda, Janssen, Abbvie, and Tarp; and consultancy fees from Takeda, Abbvie, and Taro. BC received consulting fees/speaker's honoraria from Janssen, Abbvie, Takeda, Gilead, and Novartis; and research grants from Pfizer, Ferring, and Janssen. RW received consulting fees/speaker's honoraria from AbbVie, Takeda, Janssen, and Pfizer.

Acknowledgment

We thank Anil Asthana [AA], Emma Calabrese [EC], and Rudiger Görtz [RG] for their contribution for the initial discussion on response parameters.

Author Contributions

Concept and design of the study: RW, KNo, KNy. Data acquisition: all authors. Statistical analysis and interpretation of data: RW. Preparing firtst draft: KNo, RW. Revising the work critically for important intellectual content: all authors. All authors. approved the final version of the manuscript, edited, and added intellectual content. Conference presentation. ECCO Vienna 2020, poster P176.

Supplementary Data

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

References

1. Peyrin-Biroulet L, Sandborn W, Sands BE, *et al*. Selecting Therapeutic Targets in Inflammatory Bowel Disease [STRIDE]: determining therapeutic goals for treat-to-target. *Am J Gastroenterol* 2015;**110**:1324–38.
2. Dulai PS, Jairath V. How do we treat inflammatory bowel diseases to aim for endoscopic remission? *Clin Gastroenterol Hepatol* 2020;**18**:1300–8.
3. Bryant RV, Friedman AB, Wright EK, *et al*. Gastrointestinal ultrasound in inflammatory bowel disease: an underused resource with potential paradigm-changing application. *Gut* 2018;**67**:973–85.
4. Calabrese E, Maaser C, Zorzi F, *et al*. Bowel ultrasonography in the management of Crohn's disease. a review with Recommendations of an International Panel of Experts. *Inflamm Bowel Dis* 2016;**22**:1168–83.
5. Maaser C, Sturm A, Vavricka SR, *et al*. ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohns Colitis* 2019;**13**:144–64.
6. Taylor SA, Mallett S, Bhatnagar G, *et al*.; METRIC study investigators. Diagnostic accuracy of magnetic resonance enterography and small bowel ultrasound for the extent and activity of newly diagnosed and relapsed Crohn's disease [METRIC]: a multicentre trial. *Lancet Gastroenterol Hepatol* 2018;**3**:548–58.
7. Rajagopalan A, Sathananthan D, An YK, *et al*. Gastrointestinal ultrasound in inflammatory bowel disease care: patient perceptions and impact on disease-related knowledge. *JGH Open* 2020;**4**:267–72.

8. Brenner DJ, Doll R, Goodhead DT, *et al*. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. *Proc Natl Acad Sci U S A* 2003;**100**:13761–6.
9. Jairath V, Ordas I, Zou G, *et al*. Reliability of measuring ileo-colonic disease activity in Crohn's disease by magnetic resonance enterography. *Inflamm Bowel Dis* 2018;**24**:440–9.
10. Miles A, Bhatnagar G, Hallian S, *et al*. Magnetic resonance enterography, small bowel ultrasound and colonoscopy to diagnose and stage Crohn's disease: patient acceptability and perceived burden. *Eur Radiol* 2019;**29**:1083–93.
11. Novak K, Tanyingoh D, Petersen F, *et al*. Clinic-based point of care transabdominal ultrasound for monitoring Crohn's disease: impact on clinical decision making. *J Crohns Colitis* 2015;**9**:795–801.
12. Bots S, Nylund K, Löwenberg M, Gecse K, Gilja OH, D'Haens G. Ultrasound for assessing disease activity in IBD patients: a systematic review of activity scores. *J Crohns Colitis* 2018;**12**:920–9.
13. Nylund K, Maconi G, Hollerweger A, *et al*. EFSUMB Recommendations and Guidelines for gastrointestinal ultrasound. *Ultraschall Med* 2017;**38**:e1–15.
14. Maconi G, Carsana L, Fociani P, *et al*. Small bowel stenosis in Crohn's disease: clinical, biochemical and ultrasonographic evaluation of histological features. *Aliment Pharmacol Ther* 2003;**18**:749–56.
15. Pallotta N, Vincoli G, Montesani C, *et al*. Small intestine contrast ultrasonography [SICUS] for the detection of small bowel complications in Crohn's disease: a prospective comparative study versus intraoperative findings. *Inflamm Bowel Dis* 2012;**18**:74–84.
16. Greenup AJ, Bressler B, Rosenfeld G. Medical imaging in small bowel Crohn's disease-computer tomography enterography, magnetic resonance enterography, and ultrasound: "Which One Is the Best for What?". *Inflamm Bowel Dis* 2016;**22**:1246–61.
17. Calabrese E, Kucharzik T, Maaser C, *et al*. Real-time interobserver agreement in bowel ultrasonography for diagnostic assessment in patients with Crohn's disease: an International Multicentre Study. *Inflamm Bowel Dis* 2018;**24**:2001–6.
18. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;**15**:155–63.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
20. Fraquelli M, Sarno A, Girelli C, *et al*. Reproducibility of bowel ultrasonography in the evaluation of Crohn's disease. *Dig Liver Dis* 2008;**40**:860–6.
21. Goodsall TM, Nguyen TM, Parker CE, *et al*. Systematic review: gastrointestinal ultrasound scoring indices for inflammatory bowel disease. *J Crohns Colitis* 2021;**15**:125–42.
22. Rimola J, Fernàndez-Clotet A, Capozzi N, *et al*. Pre-treatment magnetic resonance enterography findings predict the response to TNF-alpha inhibitors in Crohn's disease. *Aliment Pharmacol Ther* 2020;**52**:1563-73.
23. Kucharzik T, Wittig BM, Helwig U, *et al*.; TRUST study group. Use of intestinal ultrasound to monitor Crohn's disease activity. *Clin Gastroenterol Hepatol* 2017;**15**:535–42.e2.
24. Wilkens R, Hagemann-Madsen RH, Peters DA, *et al*. Validity of contrast-enhanced ultrasonography and dynamic contrast-enhanced MR enterography in the assessment of transmural activity and fibrosis in Crohn's disease. *J Crohns Colitis* 2018;**12**:48–56.
25. Novak KL, Kaplan GG, Panaccione R, *et al*. A simple ultrasound score for the accurate detection of inflammatory activity in Crohn's disease. *Inflamm Bowel Dis* 2017;**23**:2001–10.
26. Rimola J, Rodriguez S, García-Bosch O, *et al*. Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. *Gut* 2009;**58**:1113–20.
27. Qiu Y, Mao R, Chen BL, *et al*. Systematic review with meta-analysis: magnetic resonance enterography vs. computed tomography enterography for evaluating disease activity in small bowel Crohn's disease. *Aliment Pharmacol Ther* 2014;**40**:134–46.
28. Sævik F, Eriksen R, Eide GE, Gilja OH, Nylund K. Development and validation of a simple ultrasound activity score for Crohn's disease. *J Crohns Colitis* 2021 ;**15**:115–24.
29. Zabotti A, Filippou G, Canzoni M, *et al*. OMERACT agreement and reliability study of ultrasonographic elementary lesions in osteoarthritis of the foot. *RMD Open* 2019;**5**:e000795.
30. Kucharzik T, Wittig BM, Helwig U, *et al*.; TRUST study group. Use of intestinal ultrasound to monitor Crohn's disease activity. *Clin Gastroenterol Hepatol* 2017;**15**:535–42.e2.
31. Daperno M, Comberlato M, Bossa F, *et al*.; IGIBDEndo Group. Training programs on endoscopic scoring systems for inflammatory bowel disease lead to a significant increase in interobserver agreement among community gastroenterologists. *J Crohns Colitis* 2017;**11**:556–61.
32. Khanna R, Zou G, D'Haens G, *et al*. Reliability among central readers in the evaluation of endoscopic findings from patients with Crohn's disease. *Gut* 2016;**65**:1119–25.
33. Panés J, Feagan BG, Hussain F, Levesque BG, Travis SP. Central endoscopy reading in inflammatory bowel diseases. *J Crohns Colitis* 2016;**10**[**Suppl 2**]:S542–7.
34. Feagan B, Sandborn WJ, Rutgeerts P, *et al*. Performance of Crohn's disease clinical trial endpoints based upon different cutoffs for patient reported outcomes or endoscopic activity: analysis of EXTEND data. *Inflamm Bowel Dis* 2018;**24**:932–42.
35. Tsai R, Mintz A, Lin M, *et al*. Magnetic resonance enterography features of small bowel Crohn's disease activity: an inter-rater reliability study of small bowel active inflammation in clinical practice setting. *Br J Radiol* 2019;**92**:20180930.
36. Tielbeek JA, Makanyanga JC, Bipat S, *et al*. Grading Crohn disease activity with MRI: interobserver variability of MRI features, MRI scoring of severity, and correlation with Crohn disease endoscopic index of severity. *AJR Am J Roentgenol* 2013;**201**:1220–8.
37. Booya F, Akram S, Fletcher JG, *et al*. CT enterography and fistulizing Crohn's disease: clinical benefit and radiographic findings. *Abdom Imaging* 2009;**34**:467–75.
38. Best WR, Becktel JM, Singleton JW, Kern F Jr. Development of a Crohn's disease activity index. National Cooperative Crohn's Disease Study. *Gastroenterology* 1976;**70**:439–44.
39. 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 Thérapeutiques des Affections Inflammatoires du Tube Digestif [GETAID]. *Gut* 1989;**30**:983–9.
40. Pariente B, Cosnes J, Danese S, *et al*. Development of the Crohn's disease digestive damage score, the Lémann score. *Inflamm Bowel Dis* 2011;**17**:1415–22.