Semi-automatic assessment of the small bowel and colon in Crohn's disease patients using MRI (the VIGOR++ project)

Puylaert CAJ[1]*, Schüffler PJ[2,3]*, Naziroglu RE[4], Tielbeek JAW[1], Li Z[4], Makanyanga JC[5], Tutein Nolthenius CJ[1], Nio CY[1], Pendse DA[5], Menys A[5], Ponsioen CY[6], Atkinson D[5], Forbes A[7], Buhmann JM[2], Fuchs TJ[3], Hatzakis H[8], Van Vliet LJ[4], Stoker J[1], Taylor SA[5], Vos FM[1,4]

* These authors contributed equally to this article.

1. Department of Radiology, Academic Medical Centre, Amsterdam, the Netherlands
2. Department of Computer Sciences, Eidgenössische Technische Hochschule Zurich, Zurich, Switzerland
3. Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, United States of America
4. Department of Quantitative Imaging, Technical University Delft, Delft, the Netherlands
5. Centre for Medical Imaging, University College London Hospitals National Health Service Foundation Trust, London, England
6. Department of Gastroenterology, Academic Medical Centre, Amsterdam, the Netherlands
7. Norwich Medical School, University of East Anglia, Norwich, England
8. Biotronics3D Ltd, London, England

Corresponding author

Full name: Carl Alejandro Julien Puylaert

Postal address: Department of Radiology (Room G1-229), Meibergdreef 9, P.O 22660, 1100DD, Amsterdam, the Netherlands

E-mail: c.a.puylaert@amc.uva.nl

Telephone: 020-5662793

Keywords: Abdominal MRI, Crohn's disease, computerised image analysis, small bowel disease, colonic disease

Word count: x

ABSTRACT

Background: MRI scores show promise for evaluation of Crohn's disease (CD) activity, although reported reproducibility is variable. Potentially, reproducibility could be improved by use of computer-assisted semi-automated measurements to reduce interobserver variation. The aim of this study was to develop and validate a predictive MRI activity score for ileocolonic CD activity based on computer-assisted semi-automatic measurements of MRI features.

Methods: An MRI based disease activity score (the "VIGOR" score) was developed using a purposeful selection of both subjective radiologist observation (mural T2 signal) and semi-automatic measurements of bowel wall thickness, excess bowel wall volume and dynamic contrast enhancement (initial slope of increase; ISI) using a retrospective cohort of 27 patients with known CD against a Crohn's Disease Endoscopic Index of Severity (CDEIS) reference standard. A second, subjective score was developed based on only on radiologist observations. For validation, both scores were applied by two observer groups to a dataset of 106 patients (59 female, median age 33) with known CD prospectively recruited from two centers, who underwent consecutive MRI and ileocolonoscopy with CDEIS scoring. Three existing MRI activity scores (MaRIA, London score and CDMI) were also applied. Correlation between the five MRI scores and CDEIS was tested using Spearman rank correlation. Interobserver agreement was evaluated using the intraclass correlation coefficient (ICC).

Results: The VIGOR score (*17.1*ISI+0.2*excess volume+2.3*mural T2*), developed subjective score, MaRIA, London score and CDMI all had comparable correlation to CDEIS (Ob1/2, r=0.58/0.59, 0.39/0.51, 0.40/0.43, 0.38/0.45 and 0.34/0.48, respectively). The VIGOR score, however, had a higher ICC compared to the other activity scores (0.81 vs. 0.44–0.59). Diagnostic accuracy for a segmental CDEIS ≥3 of 80%–81% was seen for the VIGOR score, which was similar to the four other activity scores (70%–86%).

Conclusions: The new VIGOR score achieves comparable accuracy to conventional MRI activity scores, but with improved reproducibility, favoring its use for therapy evaluation and monitoring of disease activity.

What is current knowledge:

- Subjectively evaluated MRI features combined into activity scores can be used to quantify disease activity in Crohn's disease patients.
- MRI activity scores show promise for use in therapeutic evaluation and clinical trials, although varying degrees of reproducibility have been reported.

What is new here:

- The novel VIGOR score incorporates both subjectively evaluated MRI features and new semi-automatic measurements, particularly enhancement and bowel wall volume features.
- The VIGOR score showed equivalent grading accuracy, but significantly improved reproducibility compared to existing MRI activity scores.
- Using a predefined cut-off value, the VIGOR score shows good diagnostic accuracy, comparable with other activity scores.

INTRODUCTION

Crohn's disease is an inflammatory bowel disease manifesting throughout the gastrointestinal tract, although particularly affecting the small bowel and colon. Magnetic resonance imaging (MRI) is increasingly used for diagnosis and phenotyping of Crohn's disease, because it is safe, non-invasive and has high accuracy for evaluating enteric disease and extramural complications [1]. Multiple MRI features such as wall thickness and T1/T2 bowel wall signal have been validated as biomarkers of Crohn's disease activity, demonstrating good correlation with endoscopic and histopathologic grading of inflammation [2–4].

Several MRI disease activity scores have been developed and externally validated, combining multiple MRI features to predict overall disease activity [3–6]. These scores are currently slowly disseminating into clinical practice, although they are predominantly still employed as research tools. The Magnetic Resonance Index of Activity (MaRIA), for example, has been developed using the Crohn's Disease Endoscopic Index of Severity (CDEIS) as a reference standard. The MaRIA is based on quantitative measurement of relative bowel wall contrast enhancement (RCE) along with subjective evaluation of the presence of mural ulceration and abnormal T2 signal [3]. Other indices, such as the London score and Crohn's Disease MRI Index (CDMI) also rely on qualitative grading of various features by reporting radiologists [4,6]. Before such scores can be widely adopted for evaluating disease activity and therapeutic monitoring, both high accuracy across the spectrum of disease severity, *and* good reproducibility between radiologists must be proven. The current literature, however, reports variable reproducibility for many features used in MRI activity scores [6,7]. Moreover, although MRI shows high accuracy for severe disease activity (91% accuracy), diagnostic performance drops considerably for mild disease or disease in remission (62% accuracy) [8].

One potential solution to the current limitations of MRI activity scoring is to incorporate novel software solutions, which can automatically extract relevant features from the MRI data. As such, the observer variability as well as the risk of observer bias inherent to existing scores might be decreased [9]. Specifically, new MRI image processing methods are available which delineate regions of active disease based on segmentation techniques [10], providing semi-automatic measurements of bowel wall thickness and disease volume. Furthermore, software techniques have been developed which automatically extract perfusion parameters from motion corrected free-breathing dynamic contrast enhanced (DCE)-MRI [11].

We hypothesized that a scoring system combining semi-automatic software measurements with conventional subjective radiologist scoring of MRI features can improve accuracy and reproducibility in comparison to existing MRI scores. Accordingly, our aim was to develop and validate a predictive MRI score for ileocolonic CD activity incorporating novel software assisted semi-automatic measurement of MRI features using an ileocolonoscopic standard of reference, and to compare its performance with existing MRI activity scores.

METHODS

Retrospective cohort

For development of the scoring system, an independent cohort was used, consisting of 27 patients with known Crohn's disease undergoing MR enterography (MRE) and ileocolonoscopy (with segmental CDEIS scoring) within four weeks. Prior to MRE, a standardized small bowel preparation was used consisting of 4 hours fasting and 1600 mL 2.5% Mannitol solution ingested over 1 hour before the scan. This cohort was recruited for a previous study [6]. Three patients were excluded from the original cohort, because no informed consent could be obtained for future research.

Prospective cohort

Between October 2011 and September 2014, consecutive patients ≥ 18 years with suspected or known Crohn's disease and scheduled for ileocolonoscopy were recruited from two European tertiary referral centers for inflammatory bowel disease (1. Academic Medical Center (AMC), Amsterdam, the Netherlands, and 2. University College London Hospital (UCLH), London, United Kingdom). All included patients underwent MRE and ileocolonoscopy within two weeks. The Harvey-Bradshaw Index (HBI) was collected at the time of MRI [12].

Patient exclusion criteria were contraindications to MRI (e.g. pacemakers, claustrophobia), a final diagnosis other than Crohn's disease, failure to comply with the oral contrast protocol (see below), >2 weeks between MRI and ileocolonoscopy, and insufficient bowel cleansing precluding accurate mucosal assessment. Ethical permission was obtained from both institutions' medical ethics committee and written informed consent was obtained from all patients.

MRI protocol

In the prospective cohort, patients fasted for at least 4 hours before the examination and were instructed to drink a total of 2400 mL 2.5% Mannitol solution (Baxter, Utrecht, the Netherlands) split in two doses: 800 mL (3 hours prior to MRI) and 1600 mL (1 hour prior to MRI), to achieve distension of both colonic and small bowel segments. MRI examinations were performed on a 3.0 T MRI unit (Ingenia/Achieva; Philips, Best, the Netherlands) in the supine position using a phased-array body coil. The MRI protocol used in both centers is outlined in table 1. The DCE sequence consisted of 300 consecutive volumetric acquisitions at a temporal resolution of 1.2 seconds/volume. Intravenous gadolinium contrast was administered 60 seconds after

the start of the DCE sequence block using the standard contrast agent in the participating centers (Gadovist 1.0 mmol/L, Bayer Schering Pharma, Berlin, Germany; Dotarem 0.5 mmol/L, Guerbet, Paris, France). Following the DCE series, coronal and axial 3D T1-weighted spoiled gradient-echo (SPGE) images were acquired in the delayed phase (approximately 7 minutes after contrast injection). To reduce bowel peristalsis, three separate doses of 10 mg intravenous butylscopolamine bromide (Buscopan, Boehringer Ingelheim, Ingelheim, Germany) were administered during the examination. DCE images were mutually aligned using the registration method described by Li *et al.* [11,13].

Table 1 Protocol for MRI acquisition

	Plane	Slice thickness (mm)	FOV	TR (ms)	TE (ms)	Flip angle
Balanced GE	Coronal	5	380x380	2.5	1.25	60
BTFE dynamic	Coronal	10	380x380	2-2.1	1	45
T2-SSFSE	Coronal	4	380x380	628-660	60	90
T2-SSFSE	Axial	4	400x400	759	119	90
T2-w SSFSE fat saturation	Axial	7	380x380	967-1314	50	90
DCE sequence	Coronal	2.5	380x380-439	2.9	1.8	15
3D T1-w SPGE fat saturation	Coronal	2	380x380-459	2.2-2.4	1.0-1.1	10
3D T1-w SPGE fat saturation	Axial	2	380x380	2.1-2.3	1.0-1.1	10

BTFE, balanced turbo field-echo; DCE, dynamic contrast enhanced; FOV, field of view; GE, gradient echo; SPGE, spoiled gradient-echo; SSFSE, single-shot fast spin echo; TE, echo time; TR, repetition time.

Image analysis

Scans from the retrospective cohort were all individually evaluated by four observers (C.Y.N., D.P., J.S., J.M.) resulting in four evaluations per dataset [6]. MRI examinations from the prospective cohort were evaluated using online viewer software (3Dnet Suite, Biotronics3D, London, UK) by two pairs of observers (Ob1: C.Y.N, J.S.; Ob2. D.P, S.T.) with extensive experience in MRE (>1100, >800, >500 and >1500 examinations, respectively). The first pair of observers was from AMC, the second pair from UCLH. Each MRI dataset was independently evaluated by one observer from both pairs, resulting in two evaluations per dataset. Observers were

blinded to each other's findings and clinical data.

For both the retrospective and prospective cohorts, overall scan quality was first graded on a scale from 0 (non-diagnostic images) to 3 (diagnostic images without artefacts). Subsequently, the following five bowel segments were evaluated individually: the terminal ileum (most distal 20 cm of the ileum), ascending colon, transverse colon, descending/sigmoid colon and rectum. Segment distension, defined as the percentage of adequately distended bowel for diagnostic evaluation, was graded from 0 to 4 (< 20%, 20–40%, 40–60%, 60–80%, > 80%). Furthermore, all segments were evaluated regarding the MRI features included in three existing validated MRI disease activity scores (MaRIA, London and CDMI scores) (table 2). Subsequently, segmental MaRIA, London and CDMI scores were calculated for segments in the prospective dataset [3,4], as detailed in appendix 1.

Table 2 MRI features and grading categories

	Grading score			
MRI Features	**0**	**1**	**2**	**3**
London/CDMI				
Mural thickness[a]	1–3 mm	> 3–5 mm	> 5–7 mm	> 7 mm
Mural T2 signal	Equivalent to normal bowel wall	Minor increase in signal-bowel wall appears dark grey on fat saturated images	Moderate increase in signal-bowel wall appears light grey on fat saturated images	Marked increase in signal-bowel wall contains areas of white high signal approaching that of luminal content
Perimural T2 signal	Equivalent to normal mesentery	Increase in mesenteric signal but no fluid	Small fluid rim (≤ 2 mm)	Larger fluid rim (> 2 mm)
T1 enhancement	Equivalent to normal bowel wall	Minor enhancement - bowel wall signal greater than normal small bowel but significantly less than nearby vascular structures	Moderate enhancement - bowel wall signal increased but somewhat less than nearby vascular structures	Marked enhancement - bowel wall signal approaches that of nearby vascular structures
MaRIA				
Mural thickness in mm[a]				
RCE				
Edema	Absent	Present		
Ulcers	Absent	Present		

[a] Measured using electronic calipers

MRI=magnetic resonance imaging, RCE=relative contrast enhancement

Semi-automatic measurements

Using our online viewer system, the bowel's centerline was indicated on MRI by each observer by manually placing a number of widely spaced points within the lumen of the bowel through the bowel segments (see details of segment selection below) on the post-contrast coronal T1-weighted SPGE sequence (figure 1). Subsequently, the inner and outer bowel wall surfaces of the affected bowel wall were automatically delineated using the active contour segmentation method from the 3DNetSuite post-processing environment (Biotronics3D, London, United Kingdom; technical description in appendix 2) [10]. From this delineation the following automatic bowel wall thickness (ABWT) features were automatically obtained: maximum bowel wall thickness (mm), mean bowel wall thickness (mm) and excess bowel wall volume (mm^3). The excess bowel wall volume was defined as the volume of the delineated region exceeding normal thickness, calculated from the automatic mean thickness of healthy segments (no activity on MRI/endoscopy) in the retrospective cohort. Each delineation of the diseased region was also used as a 3D region of interest on DCE images to extract the initial slope of increase (ISI) of the enhancement curve [11][1].

These semi-automated measurements were performed in the following segments: (1) *all* segments in the retrospective cohort dataset (regardless of activity) (2) in *active* segments (defined as a >0 score on at least one subjective MRI feature) of the prospective cohort datasets, and (3) in *all* segments of a random subset of 50 datasets of the prospective cohort (see power calculation in statistical analysis section below).

[1] The initial slope of increase corresponds to the mathematically defined A1 feature in this reference paper.

Reference standard

For the prospective cohort, ileocolonoscopy was performed within two weeks of MRI using a standard endoscope (model CF-160L, Olympus) by either a gastroenterologist or a senior resident in gastroenterology under direct supervision of a gastroenterologist. The endoscopist applied the CDEIS to evaluate endoscopic disease [14]. The endoscopist was blinded to findings on MRI, except for cases where a balloon-dilatation procedure was indicated. In these cases, the length of stenosis on MRI was used to determine the feasibility of balloon-dilatation.

Model development

Feature selection was based on a purposeful selection process through examination of features common to previously validated MRI activity scores (MaRIA, London score and CDMI). These features usually comprise three categories: 1. mural thickness, 2. contrast enhancement (either subjectively graded or quantified using RCE) and 3. T2 mural signal intensity (contained in the MaRIA as mural edema). Although other features have been included (perimural T2 signal, ulceration), these features did not occur in any other MRI activity score and as such were not deemed essential for the new model.

Firstly, we considered all *semi-automated* models, i.e. including one semi-automatic mural thickness or volume measurement, the ISI contrast enhancement grade and one subjective T2 signal feature. A second pool of *subjective* models was formed by using one subjective feature for each of the three feature categories (i.e. not including semi-automatic features).

From both the semi-automatic and the subjective models the ‘best’ model was selected using a previously described exhaustive search method for biomarker discovery [15]. In summary, this method evaluates *all* possible combinations of MRI features as candidate models for predicting CDEIS. Specifically, the rank correlation to CDEIS of each putative model was determined in the retrospective data using a 50-fold bootstrap cross-validation [16]. Eventually, this procedure delivered two models: the top ranking *semi-automatic* model and the top ranking *subjective* model. These were termed the "VIGOR score" and the "subjective score", respectively.

Validation of MRI activity scores

The segmental VIGOR and subjective scores as well as the MaRIA, London score and CDMI scores were correlated for each observer against the segmental CDEIS reference standard in the prospective dataset (see statistical analysis section below). Segments with missing model features (i.e. due to being non-evaluable or unable to calculate semi-automatic features) were excluded, to provide a fair comparison between different models. Additionally, interobserver agreement and correlation to CDEIS were calculated for all overlapping active segments (i.e. deemed active by both observers).

Diagnostic accuracy and per-patient scores

A secondary analysis focused on diagnostic accuracy for segmental disease activity (defined as a CDEIS ≥3 [17]). Therefore, the five MRI scores were applied to bowel segments of 50 randomly selected patients (random number generation from within the complete set of complete studies). The subset sample size of 50 datasets was determined based on a power analysis using previous MRI performance data [6].

Employing an α of 0.05 and a β of 0.20, expected colonic sensitivity of 0.4 and prevalence of 0.15, expected terminal ileum specificity of 0.8 and prevalence of 0.67, the necessary number of terminal ileum and colonic segments was calculated to be 45 and 154 segments, respectively. Anticipating a segment exclusion of 10%, a total of 50 patient datasets was required.

Sensitivity, specificity, positive predictive value, negative predictive value and diagnostic accuracy were calculated based on all segments of this subset. Segmental active disease on MRI was defined for the existing activity scores using the predetermined cut-off values (MaRIA, ≥7; London score, ≥4.1; CDMI, ≥3) [3,4]. Regarding the VIGOR and subjective scores the optimal cut-off points for detection of active disease were determined using receiver-operating characteristics (ROC) analyses performed on the retrospective training data.

Per-patient MRI activity scores for both observers and global CDEIS were calculated by summing up segmental scores and dividing by the number of evaluated segments. Subsequently, a stenosis score was added to the per-patient CDEIS score if applicable [14]. For comparison, segmental scores' correlations to CDEIS and interobserver agreement were also determined in all segments of the 50-patient subset.

Statistical analysis

Previous data has shown that inclusion of healthy segments (zero-inflated data) in interobserver agreement calculation provides over optimistic estimates (). For this reason our primary analysis is based on active segments only (>0 score on at least one subjective MRI feature). Segmental and per-patient MRI scores were correlated to the segmental and global CDEIS, respectively, using Spearman rank correlation,

which were interpreted as follows: 0–0.20, very weak; ≥0.20–0.40, weak; ≥0.40–0.60, moderately; ≥0.60–0.80, strong; ≥0.80–1.00, very strong. Interobserver agreement was evaluated using weighted kappa values for ordinal data and intraclass correlation coefficients (ICC) for continuous data, using the following criteria for interpretation: 0–0.20, poor; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, good; 0.81–1.00, very good [18]. Spearman correlation coefficients were compared using the Steiger Z-test for overlapping, dependent correlations [19]. ICC's were compared by calculating the variance through bootstrapping, after which paired Student's t-tests were performed. We considered a P-value of < 0.05 to indicate a statistically significant difference. Model development and validation were implemented with R Statistical language (v3.1.2, Vienna, Austria). Descriptive statistics were analyzed using SPSS 22 for Mac (SPSS, Chicago, USA).

RESULTS

Retrospective cohort

The retrospective cohort consisted of 27 known Crohn's disease patients (127 segments evaluable by radiologist and endoscopist). Eighteen segments (6 colon, 12 rectum) were excluded from the analysis, due to severe artefacts (n=4), poor distension (n=7) and fecal residue (n=7). A further 42 segments were excluded, as semi-automatic features could not be derived in these segments for the following reasons: segment outside the DCE field-of-view (33/42), failed DCE registration (8/42) or failed segmentation (1/42). Of the 33 segments outside the DCE field-of-view, 91% were either colonic (16/33) or rectal (14/33), which was expected for this retrospective cohort, as MRI preparation and sequences were not intended for colonic evaluation.

Prospective cohort

A total of 158 patients were prospectively recruited (89 AMC, 69 UCLH). Of these, 52 patients were excluded for the following reasons: diagnosis other than Crohn's disease (n=18), > 14 days between MRI and colonoscopy (n=7), failure to comply with the oral contrast protocol (n=6), cancelled or aborted ileocolonoscopy (n=5), incomplete MRI protocol (n=14; e.g. missing sequences and incomplete imaging), insufficient bowel cleansing (n=1) and non-compliance to breathing commands due to language barrier (n=1). The final prospective study cohort consisted of 106 patients (69 AMC, 37 UCLH), for which demographics and clinical characteristics are provided in table 3. Characteristics of the 50 patient randomly determined subset used for evaluation of diagnostic accuracy and per-patient scores can be found in appendix 3. One patient experienced abdominal pain and cramping after the MRI examination, both of which were successfully treated with simple analgesia.

Table 3 Clinical characteristics of the prospective cohort

Total no. of patients	**106**
Female, n (%)	59 (56)
Age at MRI (years), median (IQR)	33 (26–44)
Previous surgery, n (%)	42 (40)
Concomitant treatments	
Anti-TNF antibodies, n (%)	30 (28)
Steroids, n. (%) of patients	18 (17)
Thiopurines, no. (%)	14 (13)
5-ASA, no. (%) of patients	19 (18)
Methotrexate, no. (%)	8 (8)
CRP (mg/L), median (IQR)	5 (1–13)
HBI, median (IQR)	5 (2–8)
CDEIS, median (IQR)	3.2 (0.5–6.4)
Montreal classification	
Age at diagnosis (years), median (IQR)	22 (17–28)
Disease location	
L1 ileal, n (%)	43 (41)
L2 colonic, n (%)	15 (14)
L3 ileocolonic, n (%)	48 (45)
L4 upper GI tract involvement, n (%)	4 (4)
Disease behavior	
B1 inflammatory	54 (51)
B2 stricturing	36 (34)
B3 penetrating	16 (15)

Perianal involvement, n (%)	23 (22)

5-ASA, 5-acetylsalicylic acid; CDEIS, Crohn's disease Endoscopic Index of Severity; CRP, C-reactive protein; GI, gastrointestinal; HBI, Harvey-Bradshaw Index; IQR, interquartile range; MRE, magnetic resonance enterography; TNF, tumor necrosis factor.

Mean scan image quality (0–3) was 2.2 (SD: 0.6). Mean distension values for terminal ileum and colon were both 3.4 (SD: 0.7). Of included segments (evaluable on MRI by the radiologist and at endoscopic intubation), Ob1 and Ob2 identified 88 and 95 active segments on MRI, respectively. In the 50-patient subset, a total of 230 and 229 segments were included for Ob1 and Ob2, respectively.

In *active* segments (>0 score on at least one subjective feature), the VIGOR score could be calculated in 83% (73/88) of segments for Ob1 and in 73% (69/95) for Ob2. In the 50-patient subset, the VIGOR score could be calculated in 73% (167/230) of segments for Ob1. When the rectum was excluded from the analysis, this rate increased to 87% (161/186). For observer 2, the VIGOR score could be evaluated in 70% (161/229) of segments, which increased to 82% (153/187) after exclusion of the rectum. Details of segment inclusion and reasons for exclusion can be found in table 4.

Table 4 Segment inclusions and exclusions

	Active segments		**Subset (n=50), all segments**		**Subset (n=50), terminal ileum and colon**	
	Ob1	*Ob2*	*Ob1*	*Ob2*	*Ob1*	*Ob2*
Total no. of segments*	**88**	**95**	**230**	**229**	**186**	**187**
Exclusions (%)	**15 (17)**	**26 (27)**	**63 (27)**	**68 (30)**	**25 (13)**	**34 (18)**
Outside DCE	3	7	42	40	12	13
Failed DCE registration	7	7	1	1	1	1
Fecal residue	3	1	6	6	2	2
Poor distension	0	2	6	6	3	3
Artefacts	0	2	0	1	0	1
Failed segmentation	2	7	8	14	7	14
Included segments (%)	**73 (83)**	**69 (73)**	**167 (73)**	**161 (70)**	**161 (87)**	**153 (82)**
Terminal ileum	54	49	39	41	39	41
Ascending colon	9	9	44	41	44	41

Transverse colon	4	2	39	38	39	38
Desc/sigmoid colon	6	9	39	33	39	33
Rectum	0	0	6	8	-	-

* All segments which could be evaluated by the radiologist and endoscopist.

Model development

The developed VIGOR and subjective models were:

$$\text{VIGOR score} = 17.1 \times \text{ISI} + 0.2 \times \text{excess volume} + 2.3 \times \text{mural T2}$$

$$\text{Subjective score} = 0.03 \times \text{RCE} + 0.9 \times \text{mural thickness (mm)} + 3 \times \text{mural T2}$$

A VIGOR score of ≥ 5.6 was determined via ROC analysis as the optimal cut-off value for active disease (CDEIS ≥3). For the subjective score, the optimal cut-off value for active disease was ≥ 4.8.

Model validation and comparison

Correlations to CDEIS for each observer pair and interobserver agreement are presented in table 5. In active segments, the VIGOR score showed moderate correlations to CDEIS (Ob1/2: r=0.58/0.59). Weak-to-moderate correlations to CDEIS were seen for the subjective score (r=0.39/0.51), MaRIA (r=0.40/0.43), the London score (r=0.38/0.45) and the CDMI (r=0.34/0.48). Significant differences were seen for Ob1 between the VIGOR score and the subjective score (p=0.04), the London score (p=0.03), the CDMI (p=0.01), but not the MaRIA (p=0.05). For Ob2, no significant differences were seen (p=0.10–0.35). The VIGOR score showed a very good ICC (0.81), while other activity scores showed moderate ICC's (0.44–0.59) (table 5). Scatter plots for all scores between observers can be found in appendix 4.

Table 5 Correlations between MRI activity scores and interobserver agreement of individual MRI features

	Observer 1 (n=73)		Observer 2 (n=69)		Interobserver agreement (n=56)	
MRI features	*r*	*p-Value*	*r*	*p-Value*	*ICC (95% CI)*	*p-Value*
VIGOR score	0.58	<0.001	0.59	<0.001	0.81 (0.56–0.91)	<0.001
Subjective score	0.39	0.001	0.51	<0.001	0.44 (0.21–0.63)	<0.001

MaRIA	0.40	0.001	0.43	<0.001	0.44 (0.21–0.63)	<0.001
London score	0.38	0.001	0.45	<0.001	0.47 (0.24–0.65)	<0.001
CDMI	0.34	0.003	0.48	<0.001	0.59 (0.40–0.74)	<0.001

MaRIA=Magnetic Resonance Index of Activity; MRI=Magnetic Resonance Imaging; VIGOR=Virtual Gastrointestinal Tract

A complete table with subset and per-patient results can be found in appendix 5.

On the 50-patient subset including all segments (active and in remission), the VIGOR score showed moderate correlation to CDEIS (Ob1/2, r=0.57/0.53) for segmental disease activity, while the correlations for the other activity scores ranged between 0.50–0.61 for Ob1 and between 0.53–0.64 for Ob2. No significant differences were seen between the VIGOR score and other activity scores for Ob1 (p=0.2–0.6). For Ob2, the CDMI and London score showed significantly higher correlation to CDEIS compared to the other activity scores (p=0.02-0.03). The VIGOR score showed a very good ICC (0.87), while other activity scores showed good to very good ICC's (0.77–0.86).

Per-patient activity scores on the 50-patient subset showed moderate correlations to CDEIS for the VIGOR score (Ob1/2, r=0.53/0.54), subjective score (r=0.60/0.57), MaRIA (r=0.58/0.51), London score (r=0.58/0.56) and CDMI (r=0.53/0.59). There were no significant differences between any pair of activity scores (p>0.05). The VIGOR per-patient scores showed a good ICC (0.77), which was not significantly different from other activity scores for which ICC's ranged between 0.71–0.79 (p>0.05) (appendix 5).

Diagnostic accuracy

The diagnostic accuracy for segmental active endoscopic disease for the five MRI scores are presented in table 6. No significant differences in diagnostic accuracy were seen between different MRI activity scores (p>0.05), except for the subjective

scores' significantly lower accuracy for Ob1 compared to other activity scores ($p<0.01$).

Table 6 Diagnostic accuracy for segmental MRI activity scores for detection of active disease (CDEIS≥3)

	Observer 1					Observer 2				
	Sensitivity	*Specificity*	*PPV*	*NPV*	*Accuracy*	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*Accuracy*
VIGOR score	76%	84%	63%	90%	81%	74%	82%	58%	90%	80%
Subjective score	78%	67%	47%	89%	70%	74%	82%	58%	90%	80%
MaRIA	67%	86%	64%	88%	81%	64%	91%	71%	88%	84%
London score	60%	96%	84%	87%	86%	57%	94%	77%	86%	84%
CDMI	60%	92%	73%	86%	83%	62%	91%	72%	87%	83%

PPV=Positive predictive value; NPV=Negative predictive value

DISCUSSION

In this development and validation study, evidence is provided for a new MRI CD activity scoring system, the “VIGOR score”, incorporating both subjective observations and semi-automatic features. The VIGOR score achieved significantly improved reproducibility in segments with active disease in comparison to existing activity scores, such as the MaRIA, London score and CDMI, as well as a new subjective score based on the best performing combination of individual MRI features of activity. The VIGOR score showed the highest correlation with the endoscopic standard of reference, although there was no consistent statistically significant difference (for both observers) in comparison with the other activity scores. In a subset of 50 patients, the VIGOR score showed similar diagnostic accuracy compared to other activity scores. When both active and inactive segments were included, correlation to CDEIS and reproducibility were higher for subjective activity scores then when considering active segments only. Simultaneously, little changes

were observed for the VIGOR score. When considering the per-patient VIGOR score, correlation with CDEIS remained moderate and reproducibility remained good.

MRI activity scores are currently being investigated for use as outcome measures in clinical trials, with some success [20,21]. Clearly, for use in multicenter studies, a high level of reproducibility between readers is imperative. Our study reports very encouraging performance characteristics for the newly developed semi-automatic model: correlation with CDEIS is at least as good as existing scores, yet interobserver agreement is considerably higher. Both wall thickness and enhancement are proven MRI biomarkers of disease activity, yet both suffer from subjectivity of measurement. By automating the process, this variability is significantly reduced. The next stage of development should now investigate the ability of the VIGOR score to monitor therapy via longitudinal studies, similar to work reported by Ordas *et al.* evaluating the MaRIA [21].

Compared to existing evaluations of MRI activity scores, we found relatively low correlations with CDEIS [5,6,21], and lower levels of sensitivity for segmental active disease. We hypothesize that this is most likely caused by the disease spectrum in our prospective cohort, with relatively high prevalence of mild disease. This is confirmed by the median CDEIS, CRP and HBI values from our prospective cohort (table 3 and appendix 3), which are much lower than those in previous studies. [3,4].

The presence of mural ulceration has been reported as very useful sign of activity and is incorporated as part of the MaRIA score. However data suggests that evaluation of ulceration on MRI is highly reader dependent [6]. For this reason, we did not include this feature in our model development. Furthermore, all five MRI scores (four of which did not include ulceration) achieved similar correlation to CDEIS and diagnostic accuracy for active segments.

In our study, we performed separate analyses on active segments and in a subset of

patients, all segments, regardless of activity. Our primary analysis was limited to active segments as large numbers of normal segments tend to skew correlative statistics and can result in over-optimistic conclusions. This is confirmed by our results; higher reproducibility was seen for subjective activity scores in the inclusive analyses of all segments (table 5 and appendix 5).

In our study we reported generalized linear regression models, since these models are easy to apply, interpretable and understandable, and take advantage of the fact that most MRI features have positive linear relationship to Crohn's disease severity. However, we also tested Tobit regression models to account for zero-inflation of the data as well as random forest models to be independent from any relationship of the data. These models did not improve our results significantly, which is why we favor the easier interpretable and applicable generalized linear regression. We applied a heuristic feature selection using important feature categories from previous studies as the basis for model development. Our feature selection used a combination of subjective *and* semi-automatic features to fully employ the capabilities of MRI for disease evaluation.

In the development and validation studies of the MaRIA score in particular, a rectal enema was used to distend the colon [3,5]. In the current study, we have shown that good image quality can also be achieved using an oral preparation with an additional 800 mL of Mannitol solution 3 hours prior to the exam; readers graded colonic distension as generally good. Such a protocol may prove more acceptable to patients by removing the need for a rectal enema. Bowel preparation for the retrospective cohort did not include specific colon preparation. Segments with poor colonic/rectal distension or fecal residue were removed from the retrospective group as these might have introduced bias to the developed model.

Our study has several limitations. The DCE sequence used in our retrospective

cohort used a smaller field of view compared to the sequence used in the prospective cohort, which limited the number of ISI data for model development. Because the field was positioned on the terminal ileum, the excluded segments were mainly colonic and rectum segments (81% of exclusions). Furthermore, in the prospective cohort, a relatively large number of rectum segments were excluded due to being out of the field-of-view on DCE. Semi-automatic software solutions together with MRI sequences continuously undergo improvement and as such, an increase in success-rate can be expected following iterations of development. However, our results do reveal the current limitations of semi-automatic features, as measurements in segments with suboptimal preparation were limited. Although subjective evaluation is also affected, human interpretation appears superior in coping with the effects of suboptimal preparation on mural thickness and contrast-enhancement.

Currently, steps are being taken to optimize the time-efficiency of semi-automatic MRI measurements and to provide full integration in viewer software. Clearly, these aspects are essential for clinical applicability, which requires easy to use techniques.

In conclusion, the use of semi-automatic features for assessment of patients with CD maintains diagnostic and grading accuracy, while improving reproducibility over conventional activity scores. This favors its use for therapy evaluation and monitoring of disease activity. Accurate and reproducible MRI scores could improve the physician's trust in these scores to make consistent and effective treatment decisions.

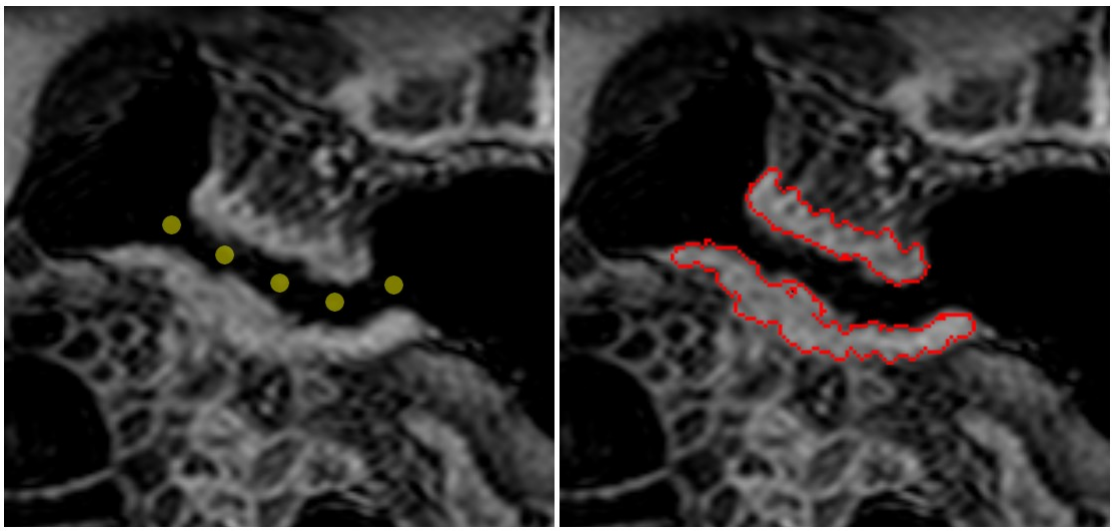

Figure 1 (A) Placement of centerline points in the lumen of an affected transverse colon segment. A few centerline points are placed in the middle of the lumen in one or more slices. (B) The delineation of the inner and outer bowel wall surfaces is visualized by a red line. Presently this is shown on a coronal slice, but it can be visualized in a similar way in reconstructed sagittal or transversal planes.

CP and PS contributed equally to this article. FV, JS, ST and LvV share senior authorship.

Contributorship statement CP, JT, JM, CYP and CT did patient recruitment; FV, JS, ST, LvV, CYP and AF designed the study; RN, ZL, LvV and FV devised the software for the semi-automatic measurements and processed measurements for this study; CYN, DP, JS and ST performed the radiologic evaluation of scans; PS, JB and TF devised and performed the exhaustive search method used for model development; DA developed the MRI protocol and provided technical MRI support throughout the project; HB is the CEO of Biotronics3D, which, as part of the VIGOR++ project, developed the platform on which MRI examinations were

evaluated and wherein the semi-automatic features were integrated; CP, JT, JM, CT and AM worked on data acquisition; CP and PS performed the analysis, interpreted the results and wrote the manuscript. All authors helped on the critical revision of the manuscript. FV is the guarantor of the article.

Acknowledgements The authors thank Ernst Harting for management of the VIGOR++ project and Costis Kompis for project exploitation, Rado Andriants imiavona and Laurence Bourne from Biotronics3D for technical support, Christopher Pawley and Asif Jaffar for patient recruitment and Isha Verkaik for database management.

Ethics approval Both clinical centers received approval from the local medical ethics committee.

Competing interests HB is CEO of the company Biotronics3D, which was a partner in the VIGOR++ project. No funding was received from Biotronics3D and HB did not have access to the data, nor was he involved in data analysis. Stuart Taylor and Jaap Stoker are MRI readers for studies in CD by Robarts Clinical Trials.

Funding The VIGOR++ project was funded through a research grant from the European Union's Seventh Framework Programme (project number 270379). The European Union was not involved in designing and conducting this study, did not have access to the data, and was not involved in data analysis or preparation of the manuscript. The project was supported by researchers at the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

Stuart Taylor is an NIHR senior investigator

Provenance and peer review Not commissioned

REFERENCES

1 Panes J, Bouhnik Y, Reinisch W, *et al.* Imaging techniques for assessment of inflammatory bowel disease: Joint ECCO and ESGAR evidence-based consensus guidelines. *J Crohn's Colitis* 2013;**7**:556–85.

2 Zappa M, Stefanescu C, Cazals-Hatem D, *et al.* Which magnetic resonance imaging findings accurately evaluate inflammation in small bowel Crohn's disease? A retrospective comparison with surgical pathologic analysis. *Inflamm Bowel Dis* 2011;**17**:984–93.

3 Rimola J, Rodriguez S, Garcia-Bosch O, *et al.* Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. *Gut* 2009;**58**:1113–20.

4 Steward MJ, Punwani S, Proctor I, *et al.* Non-perforating small bowel Crohn's disease assessed by MRI enterography: Derivation and histopathological validation of an MR-based activity index. *Eur J Radiol* 2012;**81**:2080–8.

5 Rimola J, Ordás I, Rodriguez S, *et al.* Magnetic resonance imaging for evaluation of Crohn's disease: Validation of parameters of severity and quantitative index of activity. *Inflamm Bowel Dis* 2011;**17**:1759–68.

6 Tielbeek J a W, 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. *Am J Roentgenol* 2013;**201**:1220–8.

7 Ziech MLW, Bipat S, Roelofs JJTH, *et al.* Retrospective comparison of magnetic resonance imaging features and histopathology in Crohn's disease patients. *Eur J Radiol* 2011;**80**:e299–305.

8 Horsthuis K, Bipat S, Stokkers PCF, *et al.* Magnetic resonance imaging for evaluation of disease activity in Crohn's disease: a systematic review. *Eur Radiol* 2009;**19**:1450–60.

9 Tielbeek J a W, Vos FM, Stoker J. A computer-assisted model for detection of MRI signs of Crohn's disease activity: Future or fiction? *Abdom Imaging* 2012;**37**:967–73.

10 Wang L, He L, Mishra A, *et al.* Active contours driven by local Gaussian distribution fitting energy. *Signal Processing* 2009;**89**:2435–47.

11 Li Z, Tielbeek JAW, Caan MWA, *et al.* Expiration-Phase Template-Based Motion Correction of Free-Breathing Abdominal Dynamic Contrast Enhanced MRI. *IEEE Trans Biomed Eng* 2015;**62**:1215–25.

12 Harvey RF, Bradshaw JM. A simple index of Crohn's-disease activity. *Lancet* 1980;**1**:514.

13 Li Z, Mahapatra D, Tielbeek J, *et al.* Image registration based on autocorrelation of local structure. *IEEE Trans Med Imaging* 2015;**35**:1–1.

14 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.

15 Schüffler PJ, Mahapatra D, Tielbeek JAW, *et al.* A Model Development Pipeline for Crohn's Disease Severity Assessment from Magnetic Resonance Images. *Abdom Imaging Comput Clin Appl* 2013;**8198**:1–10.

16 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. *Springer 2001* 2009;**18**:746.

17 Daperno M, Castiglione F, de Ridder L, *et al.* Results of the 2nd part Scientific Workshop of the ECCO (II): Measures and markers of prediction to achieve, detect, and monitor intestinal healing in Inflammatory Bowel Disease. J. Crohn's Colitis. 2011;**5**:484–98.

18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.

19 Steiger JH. Tests for comparing elements of a correlation matrix. Psychol. Bull. 1980;**87**:245–51.

20 Coimbra AJF, Rimola J, O'Byrne S, *et al.* Magnetic resonance enterography is feasible and reliable in multicenter clinical trials in patients with Crohn's disease, and may help select subjects with active inflammation. *Aliment Pharmacol Ther* 2016;**43**:61–72.

21 Ordás I, Rimola J, Rodríguez S, *et al.* Accuracy of magnetic resonance enterography in assessing response to therapy and mucosal healing in patients with Crohn's disease. *Gastroenterology* 2014;**146**:374–382.e1.