Semiautomatic Assessment of the Terminal Ileum and Colon in Patients with Crohn Disease Using MRI (the VIGOR++ Project)

Puylaert, Carl A.J.; Schüffler, Peter J.; Naziroglu, Robiel E.; Tielbeek, Jeroen A.W.; Li, Zhang; Makanyanga, Jesica C.; Tutein Nolthenius, Charlotte J.; Nio, C. Yung; Pendsé, Douglas A.; Menys, Alex

DOI
10.1016/j.acra.2017.12.024

Publication date
2018

Document Version
Final published version

Published in
Academic Radiology

Semiautomatic Assessment of the Terminal Ileum and Colon in Patients with Crohn Disease Using MRI (the VIGOR++ Project)

Carl A. J. Puylaert, MSc[1], Peter J. Schüffler, PhD[1], Robiel E. Naziroglu, PhD, Jeroen A. W. Tielbeek, MD, PhD, Zhang Li, PhD, Jesica C. Makanyanga, MD, Charlotte J. Tutein Nolthenius, MD, PhD, C. Yung Nio, MD, Douglas A. Pendsé, MD, PhD, Alex Menys, PhD, Cyriel Y. Ponsioen, MD, PhD, David Atkinson, PhD, Alastair Forbes, MD, PhD, Joachim M. Buhmann, PhD, Thomas J. Fuchs, PhD, Haralambos Hatzakis, Lucas J. van Vliet, PhD, Jaap Stoker, MD, PhD, Stuart A. Taylor, MD, PhD, Frans M. Vos, PhD

Rationale and Objectives: The objective of this study was to develop and validate a predictive magnetic resonance imaging (MRI) activity score for ileocolonic Crohn disease activity based on both subjective and semiautomatic MRI features.

Materials and Methods: An MRI activity score (the “virtual gastrointestinal tract [VIGOR]” score) was developed from 27 validated magnetic resonance enterography datasets, including subjective radiologist observation of mural T2 signal and semiautomatic measurements of bowel wall thickness, excess volume, and dynamic contrast enhancement (initial slope of increase). A second subjective score was developed based on only radiologist observations. For validation, two observers applied both scores and three existing scores to a prospective dataset of 106 patients (59 women, median age 33) with known Crohn disease, using the endoscopic Crohn’s Disease Endoscopic Index of Severity (CDEIS) as a reference standard.

Results: The VIGOR score (17.1 × initial slope of increase + 0.2 × excess volume + 2.3 × mural T2) and other activity scores all had comparable correlation to the CDEIS scores (observer 1: $r = 0.58$ and 0.59, and observer 2: $r = 0.34–0.40$ and 0.43–0.51, respectively). The VIGOR score, however, improved interobserver agreement compared to the other activity scores (intraclass correlation coefficient = 0.81 vs 0.44–0.59). A diagnostic accuracy of 80%–81% was seen for the VIGOR score, similar to the other scores.

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

Key Words: Crohn disease; Magnetic resonance imaging; Image interpretation, computer-assisted; Ileum; Colon.

Acad Radiol 2018; 25:1038–1045

From the Department of Radiology and Nuclear Medicine, Academic Medical Centre, Meibergdreef 9, P.O 22660, 1100DD, Amsterdam, The Netherlands (C.AJ.P., J.AW.T., C.J.T.N., C.Y.N., J.S., F.M.V.); Department of Computer Sciences, Eidgenössische Technische Hochschule Zurich, Zurich, Switzerland (P.J.S., J.M.B.); Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, New York (P.J.S., T.J.F.); Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands (R.E.N., Z.L., L.J..V., F.M.V.); College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, China (Z.L.); Center for Medical Imaging, University College London Hospitals National Health Service Foundation Trust, London, UK (J.C.M., D.A.P., A.M., D.A., S.A.T.); Department of Gastroenterology, Academic Medical Centre, Amsterdam, The Netherlands (C.Y.P.); Norwich Medical School, University of East Anglia, Norwich (A.F.); Biotronics3D Ltd, London, UK (H.H.). Received October 24, 2017; revised December 20, 2017; accepted December 25, 2017.

Sources of Funding: The VIGOR++ project was funded through a research grant from the European Union’s Seventh Framework Programme (project number 270379). The project was supported by researchers at the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Thomas Fuchs and Peter Schüffler were funded in part through the National Institutes of Health and National Cancer Institute Cancer Center Support Grant P30 CA008748.

INTRODUCTION

Crohn disease (CD) is an inflammatory bowel disease, which can present throughout the gastrointestinal tract, particularly affecting the small bowel and the colon. Magnetic resonance imaging (MRI) is increasingly used for diagnosis and phenotyping of CD because it is safe, noninvasive, and has high accuracy for evaluating enteric disease and extramural complications (1). MRI features such as wall thickness and T1 and T2 bowel wall signals have been validated as biomarkers of CD activity, demonstrating good correlation with endoscopic and histopathologic grading of inflammation (2–4). Recent years have seen several MRI disease

[1] These authors contributed equally to this article. **Address correspondence to:** C.A.J.P. e-mail: c.a.puylaert@amc.uva.nl

https://doi.org/10.1016/j.acra.2017.12.024

activity scores being developed and externally validated, combining multiple MRI features to predict overall disease activity (3–6). These scores are gradually disseminating into clinical practice, although at present, they are predominantly 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 bowel wall relative contrast enhancement, along with subjective evaluation of mural ulceration and abnormal T2 signal (3). Other indices, such as the London score and the Crohn disease MRI index (CDMI), rely on qualitative grading of various features by reporting radiologists (4,6). Such activity scores can be applied to individual bowel segments, as well as to the patient as a whole, as both are important to clinical management. Before MRI scores can be widely adopted for evaluating disease activity and therapeutic monitoring, high accuracy across the spectrum of disease severity and good reproducibility among radiologists must be proven. The current literature, however, reports variable reproducibility for many features used in MRI activity scores (6,7).

One potential solution to the current limitations of MRI activity scoring is to incorporate novel software solutions, which can automatically extract relevant features from MRI data. Such software could reduce both interobserver variability and the risk of observer bias inherent to subjective evaluation (8). New MRI image processing methods are available, which give semiautomatic measurements of bowel wall thickness, providing superior reproducibility over manual measurement (9). Further techniques have been developed that automatically extract perfusion parameters from motion corrected free-breathing dynamic contrast-enhanced (DCE) MRI (10). Although several studies have shown the potential of semiautomatic MRI assessment of CD (9–11), none of those have examined clinical practicability or validated their results using a large, independent cohort.

We hypothesize that a scoring system combining semiautomatic 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 semiautomatic measurement of MRI features using an ileocolonoscopic standard of reference, and to compare its performance with existing MRI activity scores.

MATERIALS AND METHODS

The study was divided in two phases. Firstly, a detailed modeling process was undertaken to derive two new MRI activity scores. Secondly, these new scores were validated and compared to existing scores regarding accuracy for diagnosis and grading of disease, as well as score reproducibility. Ethical permission was obtained from both institutions' medical ethics committee, and written informed consent was obtained from all patients.

Phase 1—Model Development

The modeling process employed a previously described cohort of 27 patients with known CD (6). The first developed score specifically incorporated semiautomatic measurements of bowel wall thickness and enhancement (described in more detail further in phase 2) and was termed the "virtual gastrointestinal tract (VIGOR) score." The second score incorporated only the best performing combination of a number of subjective evaluations made by radiologists (termed the "subjective score"). A full description of the model development is given in Appendix A.

Phase 2—Prospective Activity Score Testing and Model Comparison

The validation and comparison of the newly developed and existing activity scores were performed using an independent prospective cohort. Between October 2011 and September 2014, consecutive patients aged ≥18 years with suspected or known CD and scheduled for ileocolonoscopy were recruited from two European tertiary referral centres 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 MRI and ileocolonoscopy within 2 weeks. The Harvey-Bradshaw Index (HBI) was collected at the time of MRI (12).

Patient exclusion criteria were contraindications to MRI (eg, pacemakers and claustrophobia), a final diagnosis other than CD, failure to comply with the oral contrast protocol, >2 weeks between MRI and ileocolonoscopy, and incomplete MRI protocol (eg, missing sequences or incomplete imaging), or insufficient bowel cleansing precluding accurate mucosal assessment, as determined by the endoscopist.

Reference Standard

Ileocolonoscopy was performed within 2 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 (13). 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.

MRI Protocol

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 before MRI) and 1600 mL (1 hour before MRI), to achieve distension of both colonic and small bowel segments. MRI examinations were performed on a 3-T MRI unit (Ingenia and Achieva; Philips, Best, The Netherlands)

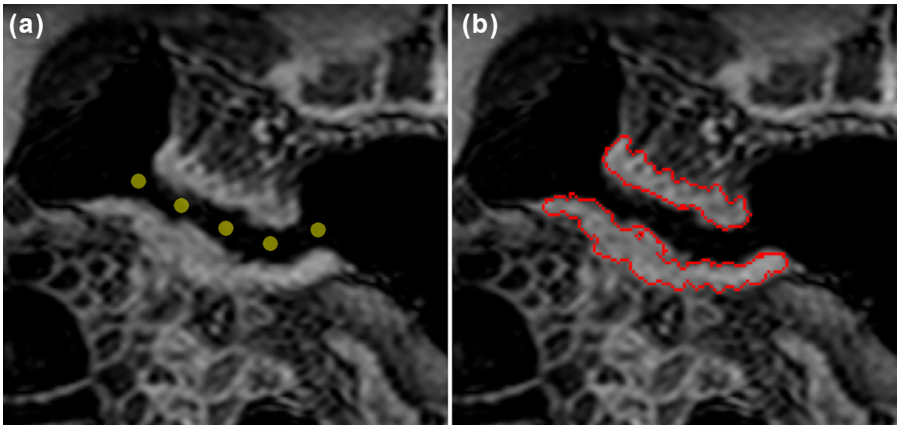

Figure 1. **(a)** Placement of centerline points in the lumen of an affected transverse colon segment on a coronal contrast-enhanced 3D T1-weighted SPGE image with fat saturation. A few centerline points are placed in the middle of the lumen in one or more slices (*yellow dots*). **(b)** The delineation of the inner and outer bowel wall surfaces is visualized by a *red lines*. Presently, this is shown on a coronal slice but can be visualized in a similar way in reconstructed sagittal or transversal planes.

in the supine position using a phased-array body coil. The MRI protocol used in both centers is outlined in Appendix A. DCE images were mutually aligned using the registration method described by Li et al. (10,14).

Image Analysis

MRI examinations 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 MR enterography (>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.

Scan quality, luminal distension, and MRI features from three existing validated MRI disease activity scores (MaRIA, London score, and CDMI) were evaluated (3,4). Details of the image analysis and the score calculation are found in Appendix A.

Semiautomatic Measurements

Using our online viewer software, the bowel's centerline was indicated on MRI individually by each observer by manually placing a number of widely spaced points within the lumen of the bowel on the postcontrast coronal T1-weighted sequence (Fig 1). If a bowel segment harbored *active* disease (defined as a >0 score on at least one subjective MRI feature), the centerline was placed across the affected part. In the absence of disease activity, the centerline was placed in a representative part of the bowel segment. Subsequently, the volume of the bowel wall was automatically delineated using the segmentation method available in our online imaging viewers' postprocessing environment (9). From this delineation, the following features were automatically obtained: maximum bowel wall thickness (mm), mean bowel wall thickness (mm), and excess bowel wall volume (mm^3) (Appendix A). Additionally, each delineation was used as a three-dimensional region of interest on DCE images to extract the initial slope of increase (ISI) of the enhancement curve (the ISI corresponds to the mathematically defined A1 feature in the reference paper) (10).

Validation of MRI Activity Scores and Statistical Analysis

Assessment of the validity of segmental scores in patients with CD can be challenging because of the high numbers of healthy segments relative to the small number of actively diseased segments (which may skew and inflate agreement statistics). For this reason, we validated the newly developed scores in two ways.

The primary validation was restricted to segments with active disease on MRI from the full prospective cohort. The applied definition of active disease (>0 score on at least one subjective MRI feature) was chosen as a low threshold to obtain the highest yield of segments in this primary analysis without creating a selection bias to one of the activity scores. The selection was not based on endoscopic disease activity, as this would require unblinding of endoscopic information to the radiologist. Grading accuracy was evaluated by correlating segmental activity scores for each observer individually against the segmental CDEIS score. Segments with missing model features (ie, nonevaluable subjective features or failure to generate semiautomatic features) were excluded, so that all activity scores were available in each segment. Additionally, interobserver agreement was calculated for all overlapping active segments (ie, deemed active by both observers) using the intraclass correlation coefficient (ICC) for absolute agreement.

The secondary validation concerned the same evaluation of grading accuracy and interobserver agreement on *all* segments (ie, active and healthy or in remission) from the subset of 50 patients. In these data, the distribution of disease forms a skewed distribution of segmental score values, violating the assumption of normality for the ICC, the standard measure for interobserver agreement in continuous data. Accordingly, we applied both the conventional ICC and a modified, nonparametric ICC by Rothery for a comprehensive evaluation of interobserver agreement (15). This measure has been

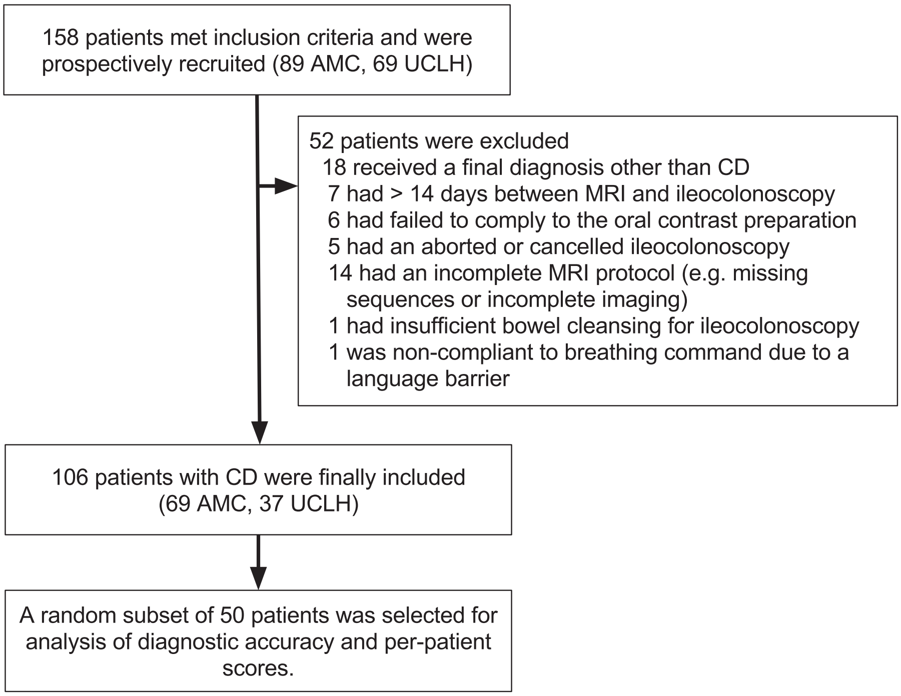
158 patients met inclusion criteria and were prospectively recruited (89 AMC, 69 UCLH)
52 patients were excluded
18 received a final diagnosis other than CD
7 had > 14 days between MRI and ileocolonoscopy
6 had failed to comply to the oral contrast preparation
5 had an aborted or cancelled ileocolonoscopy
14 had an incomplete MRI protocol (e.g. missing sequences or incomplete imaging)
1 had insufficient bowel cleansing for ileocolonoscopy
1 was non-compliant to breathing command due to a language barrier
106 patients with CD were finally included (69 AMC, 37 UCLH)
A random subset of 50 patients was selected for analysis of diagnostic accuracy and per-patient scores.

Figure 2. Flow diagram detailing patient inclusions and exclusions.

used in several studies (16,17). The subset was determined by random number generation from within the set of complete studies to minimize risk of selection bias, whereas a sample size calculation was performed using previous MRI performance data (Appendix A) (6).

In both analyses, the developed scores from phase 1 were compared to three existing MRI activity scores (MaRIA, London score, and CDMI). Diagnostic accuracy and per-patient analysis were performed using the subset of 50 patients, as detailed in Appendix A.

Spearman rank correlations were interpreted as follows: 0–0.20, very weak; ≥0–0.40, weak; ≥0.40–0.60, moderate; ≥0.60–0.80, strong; and ≥0–1.00, very strong. Correlation coefficients were then compared using the Steiger *Z* test for (non)overlapping, dependent correlations (18). Interobserver agreement (ICC or nonparametric ICC) was evaluated 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; and 0.81–1.00, very good (19). Diagnostic accuracy values were compared using the McNemar test. We considered a *P* value of <.05 to indicate a statistically significant difference. Model development and validation were implemented with R Statistical language (v3.1.2, Vienna, Austria) (20). Descriptive statistics were analyzed using SPSS 22 for Mac (SPSS, Chicago, IL).

RESULTS

Phase 1—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 receiver operating characteristic analysis as the optimal cutoff value for active disease (CDEIS score ≥3). For the subjective score, the optimal cutoff value for active disease was ≥4.8. Details of the development cohorts' segmental exclusions are shown in Appendix B.

Phase 2—Prospective Activity Score Testing and Comparison

After exclusions (Fig 2), the final prospective study cohort consisted of 106 patients with known CD, for which demographics and clinical characteristics are provided in Table 1. Characteristics of the 50 patients' randomly determined subset used for evaluation of diagnostic accuracy and per-patient scores can be found in Appendix B. One patient experienced abdominal pain and cramping after the MRI examination, which were successfully treated with simple analgesia.

The mean scan image quality (0–3) was 2.2 (standard deviation: 0.6). The mean distension value (0–4) for both terminal ileum and colon was 3.4 (standard deviation: 0.7). Within evaluable segments (evaluable on MRI by the radiologist and at endoscopic intubation), Ob1 and Ob2 identified 88 and 95 segments with active disease on MRI, respectively. In the subset of 50 patients, a total of 230 and 229 segments (both active and healthy and in remission) were evaluable 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)

TABLE 1. Clinical Characteristics of the Prospective Cohort

Total no. of patients	106
Female, *n* (%)	59 (56)
Age at MRI (y), median (IQR)	33 (26–44)
Previous surgery, *n* (%)	42 (40)
Concomitant treatments	
Anti-TNF antibodies, *n* (%)	30 (28)
Steroids, *n* (%)	18 (17)
Thiopurines, *n* (%)	14 (13)
5-ASA, *n* (%)	19 (18)
Methotrexate, *n* (%)	8 (8)
CRP (mg/L), median (IQR)	5 (1–13)
HBI value, median (IQR)	5 (2–8)
CDEIS score, median (IQR)	3.2 (0.5–6.4)
Montreal classification	
Age at diagnosis (y), 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; MRI, magnetic resonance imaging; TNF, tumor necrosis factor.

of the segments for Ob1 and in 73% (69/95) of the segments for Ob2. In the subset with 50 patients, the VIGOR score could be applied to 73% (167/230) of the segments for Ob1. Exclusion of rectum segments from the analysis increased this rate to 87% (161/186). For Ob2, the VIGOR score was applied to 70% (161/229) of the segments, which increased to 82% (153/187) after the exclusion of rectum segments. Details on the inclusion of bowel segments can be found in Table 2.

Model Validation and Comparison

Correlations to CDEIS scores for each observer pair and interobserver agreement are presented in Table 3. In *active segments*, the VIGOR score showed moderate correlations to CDEIS scores (Ob1: $r = 0.58$ and Ob2: $r = 0.59$). Weak-to-moderate correlations to CDEIS scores were seen for the subjective score ($r = 0.39$ and 0.51), the MaRIA ($r = 0.40$ and 0.43), the London score ($r = 0.38$ and 0.45), and the CDMI ($r = 0.34$ and 0.48). Significant differences were seen for Ob1 between the VIGOR score and the subjective score ($P = .04$), the London score ($P = .03$), and the CDMI ($P = .01$), but not for the MaRIA ($P = .05$). For Ob2, no significant differences were seen ($P = .10$–.35). The VIGOR score showed very good interobserver agreement in active segments (ICC = 0.81) compared to fair agreement for other activity scores (ICC = 0.44–0.59). Interobserver scatter plots for all scores can be found in Appendix B, which shows visually similar agreement for the analyses on the active segments of the full dataset and all segments of the subset, whereas in the latter, all scores show narrow clustering (ie, high reproducibility) of healthy segments.

In the subset of 50 patients including all segments (active and healthy and remission), the VIGOR score showed moderate correlation to CDEIS scores (Ob1: $r = 0.57$ and Ob2: $r = 0.53$) for segmental disease activity, whereas the correlations for the other activity scores ranged between 0.50 and 0.61 for Ob1 and between 0.53 and 0.64 for Ob2. No significant differences were seen between the VIGOR score and other activity scores for Ob1 ($P = .2$–.6). For Ob2, the CDMI and the London scores showed significantly higher correlation to CDEIS scores compared to the other activity scores ($P = .02$–.03). Conventional ICC values for active segments and all segments and nonparametric ICC values for all segments from the subset of 50 patients are shown in Table 4. It can be observed that the conventional ICC values for all segments were evidently higher compared to ICC values in active segments and the nonparametric ICC values, especially for the subjective and existing activity scores. Using the nonparametric ICC values, the VIGOR score showed very good agreement of (ICC = 0.89) compared to poor-to-fair agreement for other activity scores (ICC = 0.33–0.56), which was a significant difference ($P < .001$).

Diagnostic Accuracy

The diagnostic accuracy for all MRI scores are presented in Table 5. No significant differences in diagnostic accuracy were seen ($P > .05$), except for the subjective scores' significantly lower accuracy for Ob1 compared to other activity scores ($P < .01$).

Per-patient activity scores in the subset showed moderate correlations to CDEIS scores for the VIGOR score (Ob1: $r = 0.53$ and Ob2: $r = 0.54$), the subjective score ($r = 0.60$ and 0.57), the MaRIA ($r = 0.58$ and 0.51), the London score ($r = 0.58$ and 0.56), and the CDMI ($r = 0.53$ and 0.59). There were no significant differences between any pair of activity scores ($P > .05$). Per-patient scores showed similar (conventional) ICC values for the VIGOR score (0.77, 95% confidence interval [CI]: 0.62–0.86), the subjective score (0.71, 95% CI: 0.51–0.83), the MaRIA (0.75, 95% CI: 0.54–0.87), the London score (0.74, 95% CI: 0.57–0.84), and the CDMI (0.79, 95% CI: 0.65–0.88).

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 semiautomatic features. The VIGOR score achieved improved segmental reproducibility compared to existing activity scores, such as the MaRIA, the London score, and the CDMI. The

TABLE 2. Segment Inclusions and Exclusions

	Active Segments		Subset (*n* = 50), All Segments		Subset (*n* = 50), Rectum Excluded	
	Ob1	Ob2	Ob1	Ob2	Ob1	Ob2
Total no. of segments*	88	95	230	229	186	187
Inclusions (%)	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
Descending/sigmoid colon	6	9	39	33	39	33
Rectum	0	0	6	8	—	—
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
Artifacts	0	2	0	1	0	1
Failed segmentation	2	7	8	14	7	14

DCE, dynamic contrast enhanced; Ob1, observer 1; Ob2, observer 2.

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

TABLE 3. Correlations Between MRI Activity Scores and Crohn's Disease Endoscopic Index of Severity (CDEIS) and Interobserver Agreement in the Active Segments of the Full Prospective Cohort

	Observer 1 (*n* = 73)		Observer 2 (*n* = 69)		Interobserver Agreement (*n* = 56)
MRI Features	*r*	*P* Value	*r*	*P* Value	ICC (95% CI)
VIGOR score	0.58	<.001	0.59	<.001	0.81 (0.56–0.91)
Subjective score	0.39	.001	0.51	<.001	0.44 (0.21–0.63)
MaRIA	0.40	.001	0.43	<.001	0.44 (0.21–0.63)
London score	0.38	.001	0.45	<.001	0.47 (0.24–0.65)
CDMI	0.34	.003	0.48	<.001	0.59 (0.40–0.74)

CDMI, Crohn disease MRI index; CI, confidence interval; ICC, intraclass correlation coefficient; MaRIA, magnetic resonance index of activity; MRI, magnetic resonance imaging; VIGOR, virtual gastrointestinal tract.

TABLE 4. Interobserver Agreement for Segmental Scores of the 50-Patient Subset in Active Segments and in All Segments

	Active (*n* = 43)	All (*n* = 146)	
MRI Features	ICC (95% CI)	ICC (95% CI)	Nonparametric ICC (Rothery)
VIGOR score	0.70 (0.51–0.82)	0.87 (0.83–0.91)	0.89
Subjective score	0.44 (0.16–0.65)	0.77 (0.69–0.83)	0.53
MaRIA	0.45 (0.18–0.66)	0.77 (0.69–0.83)	0.33
London score	0.44 (0.16–0.65)	0.81 (0.75–0.86)	0.53
CDMI	0.55 (0.30–0.73)	0.86 (0.81–0.90)	0.56

CDMI, Crohn disease MRI index; CI, confidence interval; ICC, intraclass correlation coefficient; MaRIA, magnetic resonance index of activity; MRI, magnetic resonance imaging; VIGOR, virtual gastrointestinal tract.

Original ICC values are shown for both groups, whereas the nonparametric ICC is shown for all segments to account for the skewed distribution in this dataset.

VIGOR score showed similar correlation with the endoscopic standard of reference and diagnostic accuracy compared to other activity scores. The VIGOR score also showed superior performance in comparison to a new subjective score, which was developed and validated using the same cohorts. When considering the per-patient VIGOR score, correlation with CDEIS scores remained moderate and interobserver agreement remained very good. In contrast to the segmental

TABLE 5. Diagnostic Accuracy for Segmental Magnetic Resonance Imaging Activity Scores for Detection of Active Disease (Crohn's Disease Endoscopic Index [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

CDMI, Crohn disease MRI index; MaRIA, magnetic resonance index of activity; NPV, negative predictive value; PPV, positive predictive value; VIGOR, virtual gastrointestinal tract.

analyses, per-patient scores showed high agreement for all activity scores. This difference can be explained through the high reproducibility of all activity scores in healthy segments (Appendix B), which considerably influences the per-patient scores' agreement due to their high prevalence.

MRI activity scores are currently being investigated for use as outcome measures in clinical trials, with some success (21,22). Clearly, for use in multicenter studies, a high level of reproducibility between readers is imperative. Therapeutic management requires high reproducibility in both segmental and patient scores, as these serve different purposes in guidance and evaluation of surgical and medical therapies. Many patients CD have limited segmental disease (usually ileocecal disease), such that segmental reproducibility for disease activity is paramount. Conversely, a more global overview is important in those with multifocal disease. Our study reports very encouraging performance characteristics for the newly developed semiautomatic score: correlation with CDEIS scores is at least as good as existing scores, yet only the VIGOR score maintained high reproducibility in both per-segment and per-patient analyses. The next stage of development should now investigate the ability of the VIGOR score to monitor therapy via longitudinal studies, similar to the work by Ordas et al. evaluating the MaRIA (22).

Compared to existing evaluations of MRI activity scores, we found relatively low correlations with CDEIS scores (5,6,22). We hypothesize that this is caused by the disease spectrum in our prospective cohort, with relatively high prevalence of mild disease. This hypothesis is confirmed by the median CDEIS, C-reactive protein, and HBI values from our prospective cohort (Table 1 and Appendix B), which are much lower than those in previous studies (3,4). Furthermore, our results are accordant with previous results from our two inclusion centers (4,6).

The presence of mural ulceration has been reported as a useful sign of activity and is incorporated in the MaRIA. However, we did not include evaluation of ulceration in our model development as data suggest that it is highly reader dependent (6). Furthermore, all five MRI scores (four of which did not include ulceration) achieved similar correlation to CDEIS scores and diagnostic accuracy for active segments.

Our primary analysis was limited to active segments as large numbers of normal segments can skew agreement statistics and result in overoptimistic estimates. The skewing of data is confirmed by our results; increased ICC values are seen for subjective activity scores in the inclusive analyses of all segments, whereas no improved agreement is observed visually in the corresponding scatter plots or when using the nonparametric ICC values.

Our study has several limitations. The DCE sequence employed in our development cohort used a smaller field of view compared to the sequence used in the prospective cohort, which limited the amount of ISI data for model development. Because the field was positioned on the terminal ileum, the excluded segments from the development cohort were mainly colonic and rectum segments (81% of exclusions). Exclusions were improved considerably in the prospective cohort, although a relatively large number of rectum segments were excluded for being out of the field of view on DCE. Simultaneously, our results do reveal current limitations of semiautomatic features, as measurements in segments with suboptimal preparation were limited. Although subjective evaluation is also affected, human interpretation remains superior in coping with the effects of suboptimal preparation on mural thickness and contrast enhancement. However, semiautomatic software, together with MRI sequences, continuously undergoes improvement, and as such, an increase in success rate can be expected. These improvements might prove especially beneficial for inexperienced MRI readers. Although all readers in our study had extensive experience in magnetic resonance enterography, future research should explore the semiautomatic scores' application by readers of different levels of experience.

Currently, steps are being taken to further technically optimize the semiautomatic 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 semiautomatic features for the assessment of patients with CD maintains diagnostic and grading accuracy while improving reproducibility over conventional activity scores. These characteristics make it potentially suitable for therapy evaluation and monitoring of disease activity. Furthermore, accurate and reproducible MRI scores could improve the physician's trust in these scores to make consistent and effective treatment decisions.

ACKNOWLEDGMENTS

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

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 Crohns Colitis 2013; 7:556–585. http://dx.doi.org/10.1016/j.crohns.2013.02.020.
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; 17:984–993.
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; 58:1113–1120.
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; 81:2080–2088.
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; 17:1759–1768.
6. Tielbeek JAW, 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; 201:1220–1228.
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–e305. http://dx.doi.org/10.1016/j.ejrad.2010.12.075.
8. Tielbeek JAW, Vos FM, Stoker J. A computer-assisted model for detection of MRI signs of Crohn's disease activity: Future or fiction? Abdom Imaging 2012; 37:967–973.
9. Naziroglu RE, Puylaert CAJ, Tielbeek JAW, et al. Semi-automatic bowel wall thickness measurements on MR enterography in patients with Crohn's disease. Br J Radiol 2017; 20160654.
10. 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; 62:1215–1225.
11. 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.
12. Harvey RF, Bradshaw JM. A simple index of Crohn's-disease activity. Lancet 1980; 1:514.
13. 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; 30:983–989.
14. Li Z, Mahapatra D, Tielbeek J, et al. Image registration based on autocorrelation of local structure. IEEE Trans Med Imaging 2015; 35:1.
15. Rothery P. A nonparametric measure of intraclass correlation. Biometrika 1979; 66:629–639.
16. van Ierssel SH, Van Craenenbroeck EM, Conraads VM, et al. Flow cytometric detection of endothelial microparticles (EMP): effects of centrifugation and storage alter with the phenotype studied. Thromb Res 2010; 125:332–339.
17. Vuillemin A, Oppert JM, Guillemin F, et al. Self-administered questionnaire compared with interview to assess past-year physical activity. Med Sci Sports Exerc 2000; 32(July):1119–1124.
18. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull 1980; 87:245–251.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–174.
20. R Core Team. R: a language and environment for statistical computing. 2014. ISBN 3-900051-07-0. Available at: http://www.r-project.org/. R Found. Stat. Comput.
21. 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.
22. 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.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found at https://doi.org/10.1016/j.acra.2017.12.024.