Semi-automatic assessment of the terminal ileum and colon in Crohn disease patients using MRI (the VIGOR++ project)

Carl A.J. Puylaert, MSc[1]*, Peter J. Schüffler, PhD[2,3]*, Robiel E. Naziroglu, PhD[4], Jeroen A.W. Tielbeek, MD, PhD[1], Zhang Li, PhD[4,5], Jesica C. Makanyanga, MD[6], Charlotte J. Tutein Nolthenius, MD, PhD[1], C. Yung Nio, MD[1], Douglas A. Pendsé, MD, PhD[6], Alex Menys, PhD[6], Cyriel Y. Ponsioen, MD, PhD[7], David Atkinson, PhD[6], Alastair Forbes, MD, PhD[8], Joachim M. Buhmann, PhD[2], Thomas J. Fuchs, PhD[3], Haralambos Hatzakis[9], Lucas J. van Vliet, PhD[4], Jaap Stoker, MD, PhD[1], Stuart A. Taylor, MD, PhD[6], Frans M. Vos, PhD[1,4]

* These authors contributed equally to this article.

1. Department of Radiology and Nuclear Medicine, 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 Imaging Physics, Delft University of Technology, Delft, the Netherlands
5. College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, China
6. Center for Medical Imaging, University College London Hospitals National Health Service Foundation Trust, London, England
7. Department of Gastroenterology, Academic Medical Centre, Amsterdam, the Netherlands
8. Norwich Medical School, University of East Anglia, Norwich, England
9. Biotronics3D Ltd, London, England

Corresponding author
Full name: Carl Alejandro Julien Puylaert
Postal address: Department of Radiology and Nuclear Medicine,
Meibergdreef 9, P.O 22660, 1100DD, Amsterdam, the Netherlands
E-mail: c.a.puylaert@amc.uva.nl

Telephone: 020-5662793

Short title: Crohn disease: semi-automatic MRI assessment

ABSTRACT

Objectives: To develop and validate a predictive MRI activity score for ileocolonic CD activity based on both subjective and semi-automatic MRI features.

Materials and Methods: An MRI activity score (the "VIGOR" score) was developed from 27 validated MR enterography datasets including subjective radiologist observation of mural T2 signal and semi-automatic measurements of bowel wall thickness, excess volume and dynamic contrast enhancement (initial slope of increase; ISI). 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 female, median age 33) with known CD, using the endoscopic Crohn's Disease Endoscopic Index of Severity (CDEIS) as a reference standard.

Results: The VIGOR score (*17.1*ISI + 0.2*excess volume + 2.3*mural T2*), and other activity scores all had comparable correlation to CDEIS (Ob1/2, r=0.58/0.59, 0.34–0.40/0.43–0.51, respectively). The VIGOR score, however, improved interobserver agreement compared to the other activity scores (ICC=0.81 vs. 0.44–0.59). 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, favouring its use for disease monitoring and therapy evaluation.

INTRODUCTION

Crohn's disease is an inflammatory bowel disease, which can present throughout the gastrointestinal tract, 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]. 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]. Recent years have seen several MRI disease 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 relative bowel wall contrast enhancement (RCE) along with subjective evaluation of mural ulceration and abnormal T2 signal [3]. Other indices, such as the London score and Crohn's 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 between 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 as well as the risk of observer bias inherent to subjective evaluation [8]. New MRI image processing methods are available, which give semi-automatic measurements of bowel wall thickness, providing superior reproducibility over manual measurement [9]. Further techniques have been developed which automatically extract perfusion parameters from motion corrected free-breathing dynamic contrast enhanced (DCE)-MRI [10]. While several studies have shown the potential of semi-automatic MRI assessment of Crohn's disease [9–11], none of those have examined clinical practicability, nor validated their results using a large, independent cohort.

We hypothesize 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.

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 with existing scores regarding accuracy for diagnosis and grading of disease, and 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 Crohn's disease [6]. The first developed score specifically incorporated semi-automatic measurements of bowel wall thickness and enhancement (described in more detail in phase 2 below) and was termed the "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 was performed using an independent 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 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 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, >2 weeks between MRI and ileocolonoscopy, and incomplete MRI protocol (e.g. missing sequences or incomplete imaging) or insufficient bowel cleansing precluding accurate mucosal assessment, as determined by the endoscopist.

Reference standard

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 [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 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 centres 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 and CDMI scores) were evaluated [3,4]. Details of image analysis and score calculation can be found in Appendix A.

Semi-automatic measurements

Using our online viewer software, the bowel's centreline 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 post-contrast coronal T1-weighted sequence (Figure 1). If a bowel segment harboured *active* disease (defined as a >0 score on at least one subjective MRI feature), the centreline was placed

across the affected part. In the absence of disease activity it 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 the our online imaging viewers' post-processing 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 3D region of interest on DCE images to extract the initial slope of increase (ISI) of the enhancement curve (the initial slope of increase 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 Crohn's disease patients can be challenging due to 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. Segments with missing model features (i.e. non-evaluable subjective features or failure to generate semi-automatic features) were excluded, so that all activity scores were available in each segment. Additionally, interobserver agreement was calculated for all overlapping *active* segments (i.e. deemed active by both observers) using the ICC for absolute agreement.

The secondary validation concerned the same evaluation of grading accuracy and interobserver agreement on *all* segments (i.e. active and healthy/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 intraclass correlation coefficient (ICC), the standard measure for interobserver agreement in continuous data. Accordingly, we applied both the conventional ICC and a modified, non-parametric ICC by Rothery *et al.* for a comprehensive evaluation of interobserver agreement [15]. This measure has been 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, while a sample size calculation was performed using previous MRI performance data (Appendix A) [6].

In both analyses, the developed scores from phase one were compared to three existing MRI activity scores (MaRIA, London and CDMI scores). 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; ≥0.20–0.40, weak; ≥0.40–0.60, moderately; ≥0.60–0.80, strong; ≥0.80–1.00, very strong. Correlation coefficients were then compared using the Steiger Z-test for (non-)overlapping, dependent correlations [18]. Interobserver agreement (ICC or non-parametric 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; 0.81–1.00, very good [19]. Diagnostic accuracy values were compared using McNemar's test. 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) [20]. Descriptive statistics were analyzed using SPSS 22 for Mac (SPSS, Chicago, USA).

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 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. Details of the development cohorts' segmental exclusions are shown in Appendix B.

Phase 2 - Prospective activity score testing and comparison

After exclusions (Figure 2), the final prospective study cohort consisted of 106 patients with known Crohn's disease, 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.

Mean scan image quality (0–3) was 2.2 (SD: 0.6). Mean distension values (0–4) for terminal ileum and colon were both 3.4 (SD: 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/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) of segments for Ob1 and in 73% (69/95) for Ob2. In the 50-patients subset, the VIGOR score could be applied to 73% (167/230) of 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 segments, which increased to 82% (153/187) after exclusion of rectum segments. Details on inclusion of bowel segments can be found in Table 2.

Model validation and comparison

Correlations to CDEIS for each observer pair and interobserver agreement are presented in Table 3. 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 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 show visually similar agreement for the analyses on the *active* segments of the full dataset and *all* segments of the subset, while in the latter all scores show narrow clustering (i.e. high reproducibility) of healthy segments.

In the subset of 50 patients including *all* segments (active and healthy/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). Conventional ICC values for *active* segments and *all* segments and non-parametric 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 non-parametric ICC, especially for the subjective and existing activity scores. Using the non-parametric ICC, 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<0.001).

Diagnostic accuracy

The diagnostic accuracy for all MRI scores are presented in Table 5. No significant differences in diagnostic accuracy were seen (p>0.05), except for the subjective scores' significantly lower accuracy for Ob1 compared to other activity scores (p<0.01).

Per-patient activity scores in the 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). Per-patient scores showed similar (conventional) ICC's for the VIGOR score (0.77, 95%CI: 0.62–0.86), subjective score (0.71,

95%CI: 0.51–0.83), MaRIA (0.75, 95%CI: 0.54–0.87), London score (0.74, 95%CI: 0.57–0.84) and 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 semi-automatic features. The VIGOR score achieved improved segmental reproducibility compared to existing activity scores, such as the MaRIA, London score and CDMI. The 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 remained moderate and interobserver agreement remained very good. In contrast to the segmental 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 therapy. Many Crohn's disease patients 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 multi-focal disease. Our study reports very encouraging performance characteristics for the newly developed semi-automatic score: correlation with CDEIS 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 work reported by Ordas *et al.* evaluating the MaRIA [22].

Compared to existing evaluations of MRI activity scores, we found relatively low correlations with CDEIS [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 is confirmed by the median CDEIS, CRP 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 centres [4,6].

The presence of mural ulceration has been reported as a useful sign of activity and is incorporated in the MaRIA score. However, we did not include evaluation of ulceration in our model development as data suggests that it is highly reader dependent [6]. Furthermore, all five MRI scores (four of which did not include ulceration) achieved similar correlation to CDEIS 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 over-optimistic estimates. The skewing of data is confirmed by our results; increased ICC's were seen for subjective activity scores in the inclusive analyses of all segments, while no improved agreement is observed visually in the corresponding scatter plots or when using the non-parametric ICC.

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 due to being out of the field-of-view on DCE. Simultaneously, our results do reveal current limitations of semi-automatic 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, semi-automatic software together with MRI sequences continuously undergo 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 MR enterography, future research should explore the semi-automatic scores' application by readers of different levels of experience.

Currently, steps are being taken to further technically optimize the 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. 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.

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(7):556–585.

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(4):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(9):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(8):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(6):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(3):e299–e305.

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(6):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(4):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(8167):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(7):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):1–1.

15. Rothery P. A nonparametric measure of intraclass correlation. Biometrika. 1979;66(3):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(4):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(2):245–251.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174.

20. R Core Team. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. 2014:{ISBN} 3-900051-07-0, http://www.R-project.org. Available at: http://www.r-project.org/.

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(1):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(2):374–382.e1.

Figure 1

(A) Placement of centreline 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 centreline 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.

Figure 2

Flow diagram detailing patient in- and exclusions.

Appendix A: supplemental Methods

Model development

For development of the scoring systems, 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 [1]. Three patients were excluded from the original cohort, because no informed consent could be obtained for future research.

Scans from the development cohort were all individually evaluated by four observers (C.Y.N., D.P., J.S., J.M.) resulting in four evaluations per dataset [1]. All readers were unaware of the findings at the initial reading (> 1 year before study reading for all cases) and the findings from ileocolonoscopy, but were aware of patients' surgical history. MRI examinations from were evaluated using online viewer software (3Dnet Suite, Biotronics3D, London, UK). Features common to three previously validated MRI activity scores (MaRIA, London score and CDMI) were evaluated in all segments of the dataset and included in the selection process for model development. By reducing the number of features to include only the most essential, the potential validity of the developed model is increased. The included features comprised three categories: 1. mural thickness, 2. contrast enhancement (either subjectively graded or quantified using RCE) and 3. T2 mural signal intensity (classified in the MaRIA as mural edema) (see Score calculation, Appendix A). Additional

features, for example perimural T2 signal and ulceration, were not included as they are not common to all MRI activity scores. Semi-automatic measurements – maximum and mean bowel wall thickness, excess bowel wall volume and the initial slope of increase (ISI) – were calculated for all evaluated segments.

These features have been scored by four radiologists independently [1]. All samples of the four readers were used for model development, without averaging over the readers, since our model was intended to be applied on single readers' outcomes. All generalized linear regression models have been trained using R statistical language (v3.1.2) [2].

Two models were developed based on the previously mentioned three categories (mural thickness, contrast enhancement, T2 mural signal). For the first model, semi-automatic wall thickness and contrast enhancement parameters were included in the development process. For the second model, the semi-automatic measurements were excluded, relying only on subjective radiologist scores alone for mural thickness, contrast enhancement and T2 mural signal intensity (See MRI features and grading categories, Appendix A).

From both the semi-automatic and the subjective models the 'best' model was selected using a previously described exhaustive search method for biomarker discovery [3]. In summary, this method evaluated *all* possible combinations of MRI features as candidate models for predicting CDEIS, under the above constraint of having at least one feature per category. Specifically, the rank correlation to CDEIS of each putative model was determined in the retrospective data using a 50-fold bootstrap cross-validation

[4]. 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.

MRI protocol

	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.						

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 centres: gadobutrol (Gadovist 1.0 mmol/L, Bayer Schering Pharma, Berlin, Germany) or gadoterate meglumine (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.

MRI features and grading categories

MRI Features	Grading score			
	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				

Overall scan quality was 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. Luminal distension, defined as the percentage of adequately distended bowel for diagnostic evaluation, was graded for each segment from 0 to 4 (< 20%, 20–40%, 40–60%, 60–80%, > 80%).

Score calculation

Calculation of the London score, the Magnetic Resonance Index of Activity (MaRIA) and the relative contrast enhancement (RCE) using bowel wall signal intensity (SI) measured in a region of interest:

$$\text{London score} = 1.79 + 1.34 \times \text{Wall thickness} + 0.94 \times \text{mural T2 signal}$$

$$\text{CDMI} = \text{Wall thickness} + \text{T1 enhancement} + \text{mural T2 signal} + \text{perimural T2 signal}$$

$$\text{MaRIA} = 1.5 \times \text{Wall thickness (mm)} + 0.02 \times \text{RCE} + 5 \times \text{oedema} + 10 \times \text{ulceration}$$

$$\text{RCE} = \frac{\text{SI postcontrast} - \text{SI precontrast}}{\text{SI precontrast}}$$

RCE calculation did not include a noise correction term, as was used by Rimola et al [5], since inconsistent noise measurements were observed in our data, yielding arbitrary RCE values. Signal intensity values were corrected using the method described by Chenevert et al [6].

Excess bowel wall volume feature

The excess bowel wall volume was defined as the volume of the delineated region exceeding normal thickness. Normal thickness was calculated as the mean automatic thickness of healthy segments (no activity on MRI/endoscopy) in the development cohort.

Sample size calculation for subset of patients

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 were required.

Diagnostic accuracy

Diagnostic accuracy for segmental disease activity (defined as a CDEIS ≥3 [7]) was assessed by applying segmental MRI scores to *all* bowel segments of the 50 randomly selected patients. For evaluation of diagnostic accuracy, segmental disease activity on MRI was defined using these predetermined cut-off values: MaRIA, ≥7; London score, ≥4.1; CDMI, ≥3 [5,8]. For 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 development cohorts' datasets. Sensitivity, specificity, positive predictive value, negative predictive value and diagnostic accuracy were then calculated for *all* segments of the prospective subset.

Per-patient analysis

For the per-patient analysis, MRI activity scores and global CDEIS in the subset were calculated as the sum of segmental scores divided by the number of evaluated segments. A stenosis score was added to the per-patient CDEIS score if applicable [9]. Subsequently, MRI scores (per-patient and per-segment) were correlated to CDEIS and interobserver agreement was determined in *all* segments using the conventional ICC.

Appendix B: supplemental Results

Development cohorts' segment in- and exclusions

The retrospective development 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. As such, 67 segments remained.
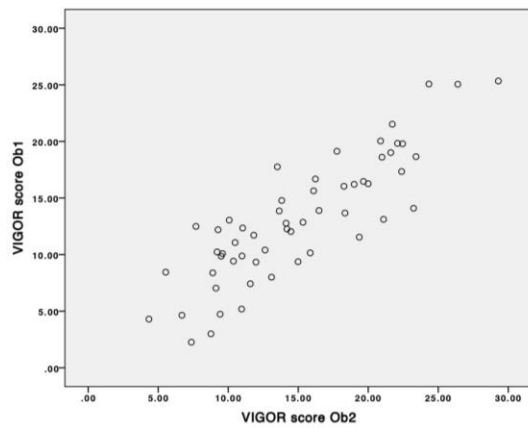
Clinical characteristics of the subset group from the prospective cohort

Total no. of patients	**50**
Female, n (%)	29 (58)
Age at MRI (years), median (IQR)	32 (27–47)
Previous surgery, n (%)	19 (38)
Concomitant treatments	
Anti-TNF antibodies, n (%)	15 (30)
Steroids, n. (%) of patients	8 (16)
Thiopurines, no. (%)	8 (16)
5-ASA, no. (%) of patients	10 (20)
Methotrexate, no. (%)	3 (6)
CRP (mg/L), median (IQR)	4 (2–11)
HBI, median (IQR)	5 (2–8)
CDEIS, median (IQR)	4.0 (0.1–7.7)
Montreal classification	
Age at diagnosis (years), median (IQR)	22 (19–28)
Disease location	
L1 ileal, n (%)	21 (42)
L2 colonic, n (%)	9 (18)
L3 ileocolonic, n (%)	20 (40)
L4 upper GI tract involvement, n (%)	2 (4)
Disease behaviour	
B1 inflammatory	25 (50)
B2 stricturing	16 (32)
B3 penetrating	9 (18)
Perianal involvement, n (%)	9 (18)

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, tumour necrosis factor.
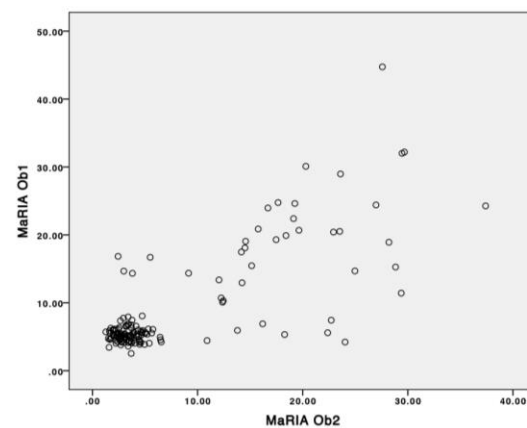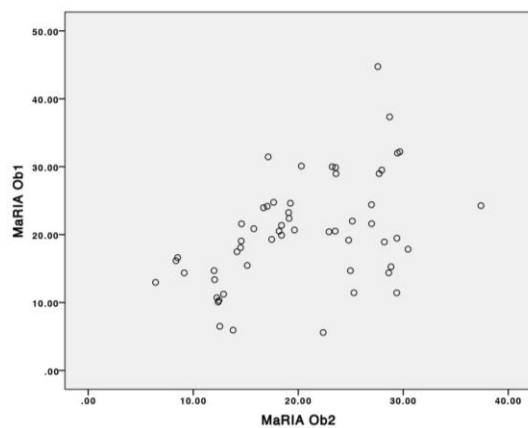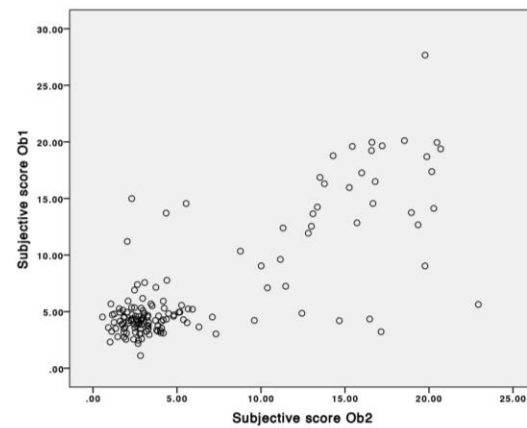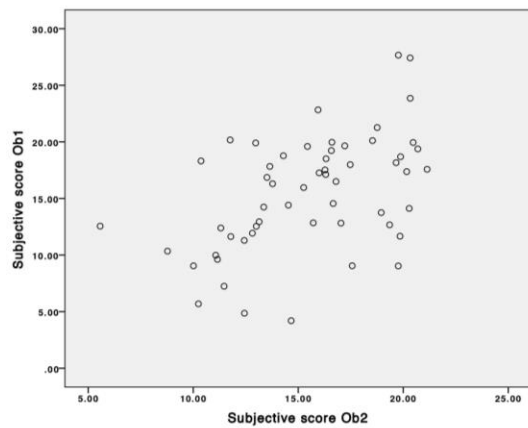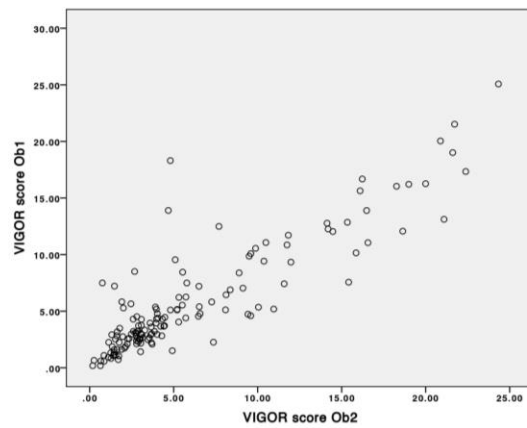
Interobserver scatter plots

Scatter plots for MRI activity scores between observer 1 (y-axis) and observer 2 (x-axis). Active (overlapping; active for both observers) segments of the full prospective cohort are shown in the left figures, while all (overlapping; included for both observers) segments (active and remission) of the 50-patient subset are shown in the figures on the right.
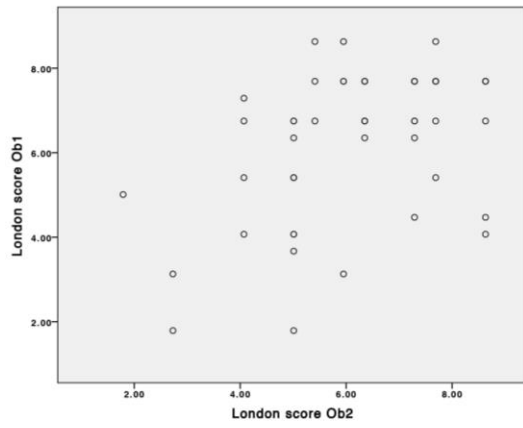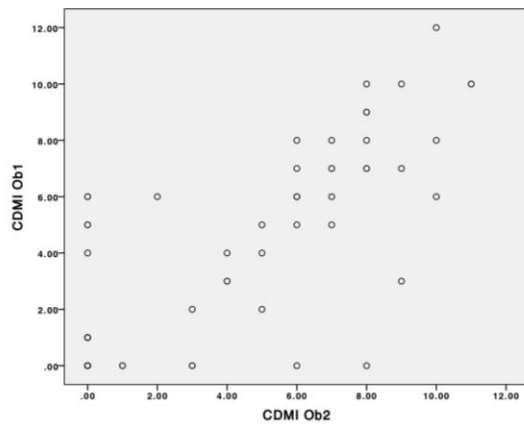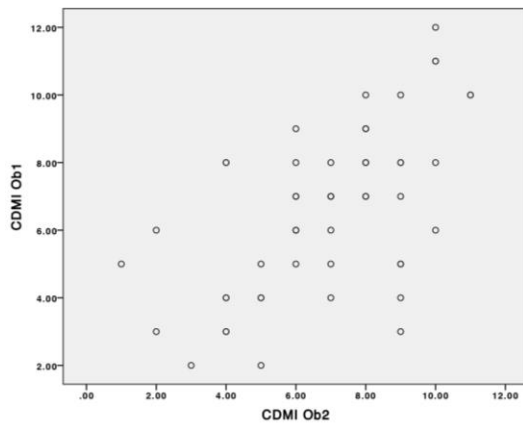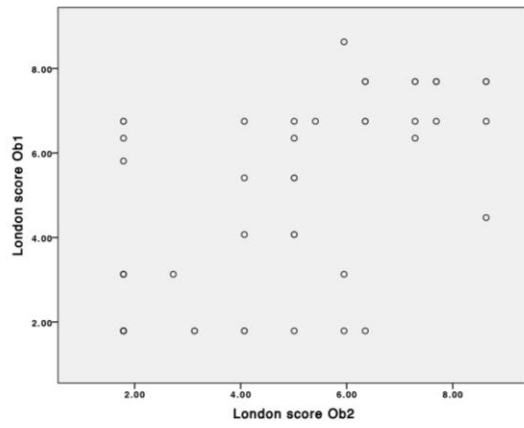
Active segments (n=56)　　　　**All segments (n=146)**

REFERENCES

1. 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(6):1220–1228.

2. R Core Team. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. 2014:{ISBN} 3-900051-07-0, http://www.R-project.org. Available at: http://www.r-project.org/.

3. 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.

4. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer 2001. 2009;18(4):746.

5. 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.

6. Chenevert TL, Malyarenko DI, Newitt D, et al. Errors in Quantitative Image Analysis due to Platform-Dependent Image Scaling. Transl. Oncol. 2014;7(1):65–71.

7. 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(5):484–498.

8. 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(9):2080–2088.

9. 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(7):983–989.

Table 1 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; MRI, magnetic resonance imaging; TNF, tumor necrosis factor.

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 segment*	**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
Desc/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
Artefacts	0	2	0	1	0	1
Failed segmentation	2	7	8	14	7	14

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

Table 3 Correlations between MRI activity scores and 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	<0.001	0.59	<0.001	0.81 (0.56–0.91)
Subjective score	0.39	0.001	0.51	<0.001	0.44 (0.21–0.63)
MaRIA	0.40	0.001	0.43	<0.001	0.44 (0.21–0.63)
London score	0.38	0.001	0.45	<0.001	0.47 (0.24–0.65)
CDMI	0.34	0.003	0.48	<0.001	0.59 (0.40–0.74)
CDMI=Crohn's Disease MRI Index; 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 *all* segments. Original ICC values are shown for both groups, while the non-parametric ICC is shown for all segments to account for the skewed distribution in this dataset.

	Active (n=43)	**All (n=146)**	
MRI features	*ICC (95% CI)*	*ICC (95% CI)*	*Non-parametric 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
MaRIA=Magnetic Resonance Index of Activity; MRI=Magnetic Resonance Imaging; VIGOR=Virtual Gastrointestinal Tract			

Table 5

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%
CDMI=Crohn's Disease MRI Index; MaRIA=Magnetic Resonance Index of Activity; PPV=Positive predictive value; NPV=Negative predictive value; VIGOR=Virtual Gastrointestinal Tract										