BJR

Received: 1 August 2016 | Revised: 10 March 2017 | Accepted: 11 April 2017

https://doi.org/10.1259/bjr.20160654

Cite this article as:
Naziroglu RE., Puylaert CA.J., Tielbeek JA.W., Makanyanga J, Menys A, Ponsioen CY., et al. Semi-automatic bowel wall thickness measurements on MR enterography in patients with Crohn's disease. *Br J Radiol* 2017; **90**: 20160654.

FULL PAPER

Semi-automatic bowel wall thickness measurements on MR enterography in patients with Crohn's disease

[1]ROBIEL E. NAZIROGLU, PhD, [2]CARL A.J. PUYLAERT, MSc, [2]JEROEN A.W. TIELBEEK, MD, PhD, [3]JESICA MAKANYANGA, MD PhD, [3]ALEX MENYS, PhD, [2]CYRIEL Y. PONSIOEN, MD, PhD, [4]HARALAMBOS HATZAKIS, MSc, [3]STUART A. TAYLOR, MD, PhD, [2]JAAP STOKER, MD, PhD, [1]LUCAS J. VAN VLIET, PhD and [1,2]FRANS M. VOS, PhD

[1]Department of Imaging Physics, Delft University of Technology, Delft, Netherlands
[2]Department of Radiology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands
[3]Center for Medical Imaging, University College London, London, UK
[4]Biotronics3D Inc., London, UK

Address correspondence to: Dr R E Naziroglu
E-mail: *R.E.Naziroglu@tudelft.nl*

Objective: To evaluate a semi-automatic method for delineation of the bowel wall and measurement of the wall thickness in patients with Crohn's disease.
Methods: 53 patients with suspected or proven Crohn's disease were selected. Two radiologists independently supervised the delineation of regions with active Crohn's disease on MRI, yielding manual annotations (Ano1, Ano2). Three observers manually measured the maximal bowel wall thickness of each annotated segment. An active contour segmentation approach semi-automatically delineated the bowel wall. For each active region, two segmentations (Seg1, Seg2) were obtained by independent observers, in which the maximum wall thickness was automatically determined. The overlap between (Seg1, Seg2) was compared with the overlap of (Ano1, Ano2) using Wilcoxon's signed rank test. The corresponding variances were compared using the Brown–Forsythe test. The variance of the semi-automatic thickness measurements was compared with the overall variance of manual measurements through an F-test. Furthermore, the intraclass correlation coefficient (ICC) of semi-automatic thickness measurements was compared with the ICC of manual measurements through a likelihood-ratio test.
Results: Patient demographics: median age, 30 years; interquartile range, 25–38 years; 33 females. The median overlap of the semi-automatic segmentations (Seg1 *vs* Seg2: 0.89) was significantly larger than the median overlap of the manual annotations (Ano1 *vs* Ano2: 0.72); $p = 1.4 \times 10^{-5}$. The variance in overlap of the semi-automatic segmentations was significantly smaller than the variance in overlap of the manual annotations ($p = 1.1 \times 10^{-9}$). The variance of the semi-automated measurements (0.46 mm^2) was significantly smaller than the variance of the manual measurements (2.90 mm^2, $p = 1.1 \times 10^{-7}$). The ICC of semi-automatic measurement (0.88) was significantly higher than the ICC of manual measurement (0.45); $p = 0.005$.
Conclusion: The semi-automatic technique facilitates reproducible delineation of regions with active Crohn's disease. The semi-automatic thickness measurement sustains significantly improved interobserver agreement.
Advances in knowledge: Automation of bowel wall thickness measurements strongly increases reproducibility of these measurements, which are commonly used in MRI scoring systems of Crohn's disease activity.

INTRODUCTION

MRI can be used for diagnostic purposes and for monitoring bowel inflammation in patients with Crohn's disease.[1–3] In particular, multiple MRI features, such as mural contrast enhancement, T_2 signal and bowel wall thickness, are useful for assessment of Crohn's disease activity.[4–7] Several disease-grading systems using these features have been validated against endoscopy and/or histopathology. Notable examples include the Magnetic Resonance Index of Activity, Clermont, London and Crohn's Disease MRI Index scores.[4,6,8–10] An important feature that is included in all the aforementioned disease activity scores is the maximal bowel wall thickness of a diseased segment.

Bowel wall thickening results from transmural inflammatory processes, including accumulation of inflammatory cells and submucosal oedema, and, in later stages, due to the formation of fibrosis. In clinical practice, the bowel wall thickness in active Crohn's disease is manually measured using electronic calipers in the most thickened part of a bowel segment. These manual measurements have limited accuracy and reproducibility due to measuring difficulties, such as partial volume effects, the complicated three-dimensional geometry of bowel loops and variable perception of the most thickened part. Accordingly, varying interobserver agreement has been reported, reflecting the subjectivity of the technique.[7,11]

Automation of the thickness measurement might improve the accuracy and reproducibility. Simultaneously, the reproducibility of the Crohn's disease scores could be enhanced.

A potential approach to perform the automated measurement might rely on (semi-)automatic delineation of the inner and outer surfaces of the bowel wall ("segmentation"). A common image-processing method for segmentation is the so-called active contour technique.[12–15] Here, a very coarse outline of an object is initially created, which is subsequently refined to the actual outline of the object of interest. Although this technique was frequently used for segmentation (*e.g.* by Lynch et al[16] and Wang et al[17]), it is not directly applicable to delineation of the bowel wall in MRI. Particularly, the inhomogeneity of the MRI signal and the thinness of the structures make segmentation a challenging problem. To our knowledge, only one automatic method for measurement of bowel thickness in MRI images has been (very recently) described.[18] This article takes a sophisticated approach to semi-automatically measure the thickness at every point along segments of the bowel. As such, it does not provide a single measure representing the segmental bowel wall thickness of which the agreement can be directly compared with that of manual measurement. Instead, the article compares observer measurements to the closest automated thickness measurements. These measurements concerned evenly distributed locations along the large bowel of patients diagnosed with Crohn's disease and not particularly regions showing active disease.

We have developed an active contour approach to semi-automatically segment both the bowel wall's inner and outer surfaces taking into account the inhomogeneous bowel content[15] (Appendix B). Subsequently, the single largest distance between the two surfaces is determined to yield the thickness measure. The method is integrated in the 3DNetSuite image post-processing environment (Biotronics3D Inc., London, UK). As such, our method is one of the first algorithms for semi-automated wall thickness measurement in MRI.

The aim of the current study is to determine the reproducibility of this semi-automatic method to measure wall thickness, especially in regions displaying active Crohn's disease, in comparison to manual measurements. We hypothesize that the semi-automatic method significantly improves interobserver agreement compared with manual measurement.

METHODS AND MATERIALS

Data

The data employed in this article were taken from two studies on Crohn's disease: (1) data from a prior, single-centre study referred to as retrospective data;[19] (2) data from a recently concluded multicentre study referred to as prospective data (publication in preparation). The local Medical Ethics Committee (1. the Academic Medical Center (AMC), Amsterdam, the Netherlands and 2. The University College Hospitals (UCLH), London, United Kingdom) approved both studies. All patients gave written informed consent to usage of their data for future investigations.

Inclusion criteria for the retrospective study were: patients with histologically proven Crohn's disease, 18 years of age or older, undergoing MRI and ileocolonoscopy (within 2 weeks) as part of their clinical follow-up in a single tertiary centre (the Academic Medical Center, Amsterdam, the Netherlands). Inclusion criteria for the prospective study were: patients with suspected or proven Crohn's disease (based on clinical data, endoscopy or histopathology), 18 years of age or older and undergoing MRI and ileocolonoscopy within 2 weeks as part of their clinical follow-up in one of two tertiary centres (1. the Academic Medical Center, Amsterdam, the Netherlands or 2. The University College Hospitals, London, United Kingdom). Exclusion criteria for both studies were: general contraindications for MRI (claustrophobia, pregnancy, renal insufficiency and pacemaker), an incomplete scan protocol or incomplete colonoscopy, *e.g.* due to impassable strictures.

The retrospective data were from all 27 patients consecutively included in the prior study at AMC between February 2009 and November 2010. Patients drank 1600 ml of a hyperosmolar fluid (mannitol, 2.5%; Baxter, Utrecht, Netherlands) 1 h before acquiring the MRI scans for optimal distension of the terminal ileum. MRI was performed on a 3.0-T MRI scanner (Intera; Philips Healthcare, Best, Netherlands). Imaging included amongst others a breath-hold contrast-enhanced T_1 weighted spoiled gradient echo series with fat saturation. This was the sequence that was used for semi-automatic thickness measurement (see section Semi-automatic thickness measurement).

The prospective data were from 26 patients randomly selected from the prospective study data acquired at UCLH. This was performed to have an approximately equal number of patients from two different medical centres. Patients were consecutively included in the prospective study from December 2011 until August 2014. The UCLH data were acquired with almost the same imaging protocol as the retrospective data. The most relevant difference concerned the patient preparation, which involved an additional ingestion of 800 ml of mannitol (2.5%) 3–6 h prior to the examination to optimize the distension of the colon. MRI was also performed on a 3.0-T MRI scanner (Achieva®; Philips Healthcare).

Detailed scan protocols are listed in Appendix A.

Annotations

Two experienced abdominal radiologists (JS (>800 enterographies, 21 years' experience) and ST (>1600 enterographies, 13 years' experience)) independently identified all regions they considered to represent active Crohn's disease. The presence of active Crohn's disease was based on all available MRI sequences. Two research fellows (JT and AM, respectively) independently annotated the data on behalf of the radiologists. Henceforth, these annotations are referred to as Ano1 and Ano2, respectively. Specifically, each annotation was performed by successively drawing (two-dimensional) polygons in all slices including the diseased segment on the coronal contrast-enhanced T_1 image. The stacks of two-dimensional polygons constituted three-dimensional volumes of active disease segments and served as references for the semi-automatic segmentations (see below). Annotations were considered to correspond (*i.e.* Ano1 and Ano2, detailing the same diseased segment) if they had at least 10% overlap.

The same two radiologists (JS and ST) and one research fellow (CP (>100 enterographies)) manually measured the maximal

bowel wall thickness of each segment with disease activity, *i.e.* with an annotation Ano1 or Ano2, including segments with overlap between Ano1 and Ano2. The three observers were instructed to measure the bowel wall thickness of each segment, but no further guidance or instruction was given. In particular, observers were BLINDED to the measurements of each other. Henceforth, these measurements will be referred to as Ob1, Ob2 and Ob3.

Semi-automatic thickness measurement

The method to measure the bowel wall thickness comprised four main steps:

(1) initialization
(2) identification of the bowel wall's inner surface
(3) identification of the bowel wall's outer surface
(4) thickness measurement.

The method was initialized by manually placing a few points to indicate the bowel's centerline across a diseased section of the gastrointestinal tract (Figure 1a). Subsequently, a small, virtual tube was constructed around this centerline that served as an initial model for the bowel wall's inner surface.

Next, the initial model was mathematically deformed, such that it delineated the transition from the bowel lumen to bowel wall as inferred from the MRI data. For that purpose, we applied a well-described technique in image processing, referred to as active contour segmentation.[12–15] The technique takes into account the inhomogeneous content of the bowel lumen (Appendix B).

Subsequently, we fixed the delineated bowel wall's inner surface and applied a similar active contour segmentation to identify the wall's outer surface. This second segmentation was initialized by outwardly dilating the inner surface segmentation by 4 mm, slightly larger than the average thickness of the healthy bowel wall.[20] This initial model was also mathematically deformed such that it delineated the transition from the bowel wall to the adjacent tissues. The method was very similar to the one identifying the bowel wall's inner surface except for considering structures with varying signal intensity outside the bowel wall.

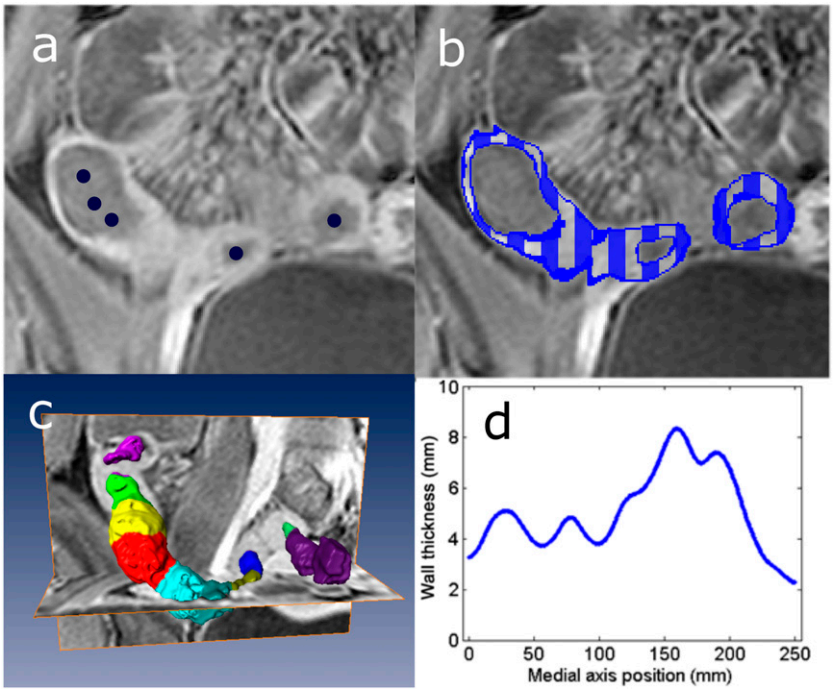
Figure 1b shows an example of the resulting segmentation of the bowel wall.

Finally, the distance was measured from each point on the segmented inner surface to the closest point on the segmented outer surface. These distances were averaged over small patches of 4 mm in length along the circumference of the bowel. The maximum average value was taken as the wall's maximal local thickness. Figure 1c,d illustrates this procedure.

The mathematical details of the method are described in Appendix B.

For the current study, in order to initiate the algorithm, two research fellows (RN, CP) independently indicated a centerline through the diseased segments delineated by annotations Ano1 and Ano2. This was carried out to allow subsequent comparison between the manual annotations and the semi-automatic segmentations.

Figure 1. (a) Initialization by placing a few points (dark blue) on the bowel's centerline. (b) Example segmentation of the bowel wall indicated by a dark blue line and a striped pattern. (c) The average thickness determined over small patches along the circumference of the bowel; separate colours indicate such patches. (d) Average patch thickness as a function of the position on the centerline. The maximum value represented the wall's maximal thickness, *i.e.* 8.2 mm in the example.



Centerlines were drawn in the bowel lumen to completely cross the full extent of the annotated disease segments. Subsequently, the described algorithm yielded segmentations (Seg1 for research fellow 1 and Seg2 for research fellow 2) and respective measures of the maximum bowel wall thickness (M1 and M2).

Using this methodology, annotated regions Ano1 corresponded to semi-automatic segmentations Seg1 and the annotated regions Ano2 to the semi-automatic segmentations Seg2. Furthermore, in segments with overlapping Ano1 and Ano2 annotations, the manual measurements Ob1, Ob2 and Ob3 corresponded to measurements (M1 and M2). In annotated segments without overlap, Ob1, Ob2 and Ob3 corresponded with only one measurement (either M1 or M2).

Evaluation measures

The performance of the semi-automatic segmentation procedure and the subsequent thickness measurement were separately evaluated by:

- quantifying the overlap (correspondence) between the semi-automatic segmentations and manual annotations
- assessing the distance (*i.e.* the mismatch of the contours) between the semi-automatic segmentations and manual annotations
- visually grading the overlap (correspondence) between the semi-automatic segmentations and manual annotations
- correlating the semi-automatic thickness measurements to the manual measurements.

A coefficient was determined reflecting the overlap or correspondence between the manual annotations and the semi-automatic segmentations (Figure 2a). In all cases, a semi-automatic segmentation covered a larger part of the bowel wall than the corresponding manual annotation. This was because the centerlines drawn by the research fellows extended beyond the manual annotations for some distance. The overlap coefficient was calculated as the percentage of volume of the manual annotations that did not overlap with the semi-automatic segmentation. This overlap measure is referred to as the semi-Dice coefficient, as it is essentially an asymmetric version of the Dice coefficient that is often applied in image processing research to measure overlap.[21]

Furthermore, the mean shortest distance was determined from points on the manual delineations Ano1 and Ano2 to the semi-automatic segmentations Seg1 and Seg2 (Figure 2b).

The research fellows who drew the paths also visually graded the accuracy of the semi-automatic segmentation independently on a four-point Likert scale. The Likert scale reflected the percentage of overlap with the perceived lesion: 0: 0–50%, no overlap; 1: 51–70%, poor overlap; 2: 71–90%, moderate overlap; 3: 91–100%, complete overlap. We opted to apply this skewed scale as any overlap $<$50% essentially represents a failed segmentation.

Finally, the intraclass correlation coefficient (ICC) between the manual thickness measurements Ob1, Ob2 and Ob3 and the thickness measurements derived from the semi-automatic segmentations was determined. ICC values were interpreted using the following criteria: 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.[22]

Statistical analysis

The overlap of the semi-automatic segmentations was statistically compared with the overlap of the manual annotations (Ano1 and Ano2) using the Wilcoxon's signed-rank test. The associated variances were statistically assessed by the Brown–Forsythe test. The same statistical tests were used to compare the mean shortest distances of the segmentations to the annotations.

The variance of the automated measurements was compared with overall variance of the manual measurements through an F-test. The ICCs of the manual thickness measurements were statistically compared with the semi-automatic measurements by means of a generalized Bland–Altman procedure. The overall ICC of the (three) manual measurements was statistically compared with the ICC of the (two) semi-automatic measurements through a likelihood-ratio test using a standard mixed model analysis.

A value of $p < 0.05$ was considered statistically significant. All statistical analyses were performed with IBM SPSS® Statistics v. 22.0 for Microsoft® Windows® computers (IBM Corp., New York, NY; formerly SPSS Inc., Chicago, IL).

Figure 2. (a) The semi-Dice coefficient is defined as the volume of the annotation (A) not covered by the segmentation (S), *i.e.* the ratio of the striped volume to the dotted plus striped volume of the annotation. (b) The mean shortest distance between annotation (A) and segmentation (S) is calculated by sampling points on A and averaging the distance of each such point to the closest position on S.

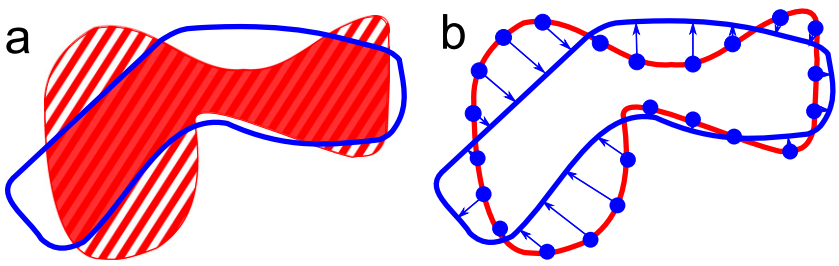
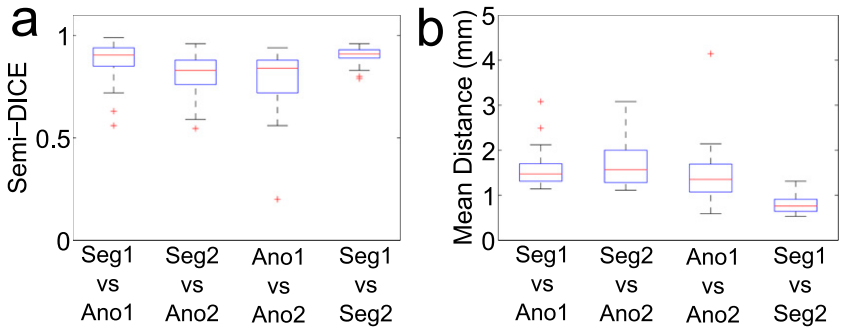


Figure 3. Comparison of the manual annotations Ano1 and Ano2 with the semi-automatic segmentations Seg1 and Seg2. Only corresponding regions are included, *i.e.* in which there was at least 10% overlap between the annotations Ano1 and Ano2 ($n = 30$). Each box plot shows the distribution of semi-Dice coefficients (a) and the mean shortest distances (b), respectively, for a particular comparison (horizontally). The boxes display the median, 25th and 75th percentiles, respectively, of the data distribution; whiskers extend to 1.5 times the interquartile range; and values outside these ranges are indicated as individual points.



RESULTS

The demographics of included patients were: median age, 30 years; interquartile range, [25, 38]; 33/53 females. In the complete data set, 52/53 patients were identified with active Crohn's disease on MRI by either one of the radiologists or by both. In these patients, there were 47 Ano1 annotations and 42 Ano2 annotations. Across both radiologists, there were 59 unique segments identified as active (*i.e.* annotated either as Ano1 or Ano2 alone, or by both), of which 30 corresponded, *i.e.* by having an overlap of >10%. The median overlap of these corresponding regions was 72%.

Evaluation of the semi-automatic segmentations

In Figure 3, the manual annotations (Ano1 and Ano2) in the corresponding segments (n = 30) are compared with the semi-automatic segmentations Seg1 and Seg2. Figure 3a shows distributions of the semi-Dice coefficient (*i.e.* the overlap measure) and Figure 3b shows the distributions of the mean shortest distances.

The median semi-Dice coefficients were: Ano1 *vs* Seg1 = 0.87; Ano2 *vs* Seg2 = 0.76; Ano1 *vs* Ano2 = 0.72; and Seg1 *vs* Seg2 = 0.89.

The overlap of the semi-automatic segmentations (Seg1 *vs* Seg2) was significantly greater than the overlap of the two manual annotations (Ano1 *vs* Ano2): $p = 1.4 \times 10^{-5}$. Also, the variance in the overlap of the semi-automatic segmentations was significantly smaller than the variance in the overlap of the manual annotations ($p = 1.1 \times 10^{-9}$).

The median of the mean shortest distances were: Ano1 *vs* Seg1 = 1.31; Ano2 *vs* Seg2 = 1.28; Ano1 *vs* Ano2 = 1.07; and Seg1 *vs* Seg2 = 0.64.

The median of the distances between the semi-automatic segmentations (Seg1 *vs* Seg2) was significantly smaller than the median of the mean shortest distance between the two manual annotations (Ano1 *vs* Ano2); $p = 6.0 \times 10^{-6}$. Also, the variance in the mean shortest distance of the semi-automatic segmentations was significantly smaller than the variance in the mean shortest distance of the manual annotations ($p = 1.5 \times 10^{-9}$).

Figure 4 is a bar chart summarizing the visual assessment of the semi-automatic segmentations, for all annotated regions. The distribution over the grades (0–3) for Seg1 was: 2, 1, 10 and 34, respectively ($n = 47$). The distribution for Seg2 over the grades was: 1, 1, 4 and 36, respectively ($n = 42$). The segmentations with grade 0 and grade 1 related to images with imaging artefacts ($n = 1$) and extensive faecal residue obscuring the bowel wall ($n = 2$), respectively.

Figure 5 relates box plots of the semi-Dice coefficients of the segmentations to the grades given for all segmentations: Seg1 (a) and Seg2 (b), respectively. Additionally, Table 1 shows the medians of the semi-Dice coefficients, mean shortest distances and visual grades for all segmentations.

Evaluation of the semi-automatic thickness measurement

Table 2 details the paired ICCs of the wall thickness measurements of the corresponding regions: Obs1, Obs2, Obs3, M1 and

Figure 4. Bar chart showing the visual assessment of Seg1 and Seg2 by the research fellows initiating the segmentations. Horizontally are the Likert gradings reflecting overlap of the segmentations with the annotations: 0: 0–50%, no overlap; 1: 51–70%, poor overlap; 2: 71–90%, moderate overlap; and 3: 91–100%, complete overlap. Vertical are the fraction of segmentations in a grading category (summing to 1 for both fellows). All annotations are included ($n = 47$ for Seg1; $n = 42$ for Seg2).

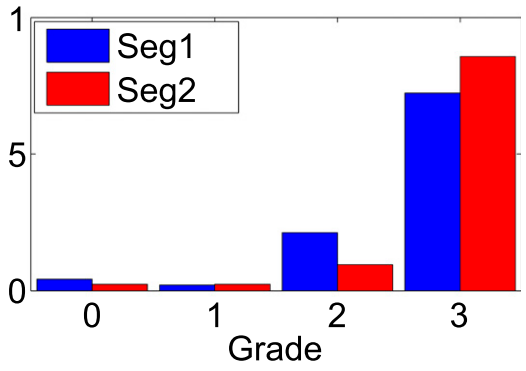
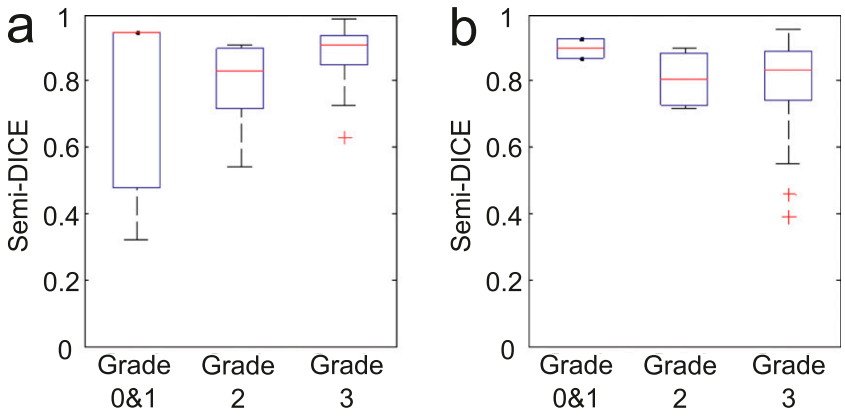


Figure 5. Box plots of semi-Dice coefficients as a function of the visual grades. (a) Seg1 ($n = 3$ for Grades 0 and 1; $n = 10$ for Grade 2; and $n = 34$ for Grade 3); (b) Seg2 ($n = 2$ for Grades 0 and 1; $n = 4$ for Grade 2; and $n = 36$ for Grade 3).



M2. Similarly, Tables 3 and 4 show the ICCs of the wall thickness measurements of all regions, *i.e.* depicting the ICCs for M1 and M2, respectively, separately. There is one semi-automatic measurement in the latter tables because there is no corresponding semi-automatic measurement for some regions.

The ICCs of the paired manual measurements of the overlapping segments varied from fair [lowest ICC: 0.34 (Obs1 *vs* Obs2)] to good [highest ICC: 0.60 (Obs2 *vs* Obs3)]. The overall ICC of the manual measurement was moderate: 0.45. The ICC of the semi-automatic measurements of these segments was considered very good (0.88).

Figure 6 visualizes the variation of the manual and the automated measurements on the corresponding segments. The overall variance of the manual measurements was 2.9 mm^2; the variance of semi-automated measurement was significantly smaller: 0.46 mm^2 ($p = 1.1 \times 10^{-7}$).

The ICC of the semi-automatic measurements was significantly higher than each ICC of the manual measurements. Particularly, $p = 7.7 \times 10^{-5}$ for the comparison of the ICC of Obs2 *vs* Obs3 (0.60) to the ICC of M1 *vs* M2. Moreover, the overall ICC of the semi-automatic measurements was found to be significantly lower than the overall ICC of the manual measurement ($p = 0.0054$).

DISCUSSION

This study evaluated a semi-automatic method to measure bowel wall thickness. The method consisted of four steps: (1) initialization by manually drawing a centerline, (2) segmentation of the bowel wall's inner surface, (3) segmentation of the wall's outer surface and (4) measurement of the bowel wall thickness.

The overlap and distance between semi-automatic segmentations and manual annotations were first determined on the corresponding annotations (*i.e.* where the two independent manual annotations were overlapping for >10%). This allowed a direct comparison of the performance measures. The median overlap of manual and automatic segmentations with these annotations was large: 0.87 (Seg1 *vs* Ano1) and 0.76 (Seg2 *vs* Ano2). Additionally, the median distance between segmentation and manual annotations was small: 1.31 mm (Seg1 *vs* Ano1) and 1.28 mm (Seg2 *vs* Ano2). What is more, the two semi-automatic segmentations had significantly larger overlap and shorter distance to each other than the manual annotations. There was also a significantly smaller spread in overlap and distance between the semi-automatic segmentations than between the annotations. This signifies the good reproducibility of the semi-automatic segmentations.

The corresponding segments might be relatively "easy" to segment because there was agreement between the annotators regarding the presence of active disease. However, the large majority of visual gradings indicated complete overlap of segmentation and annotation considering all regions (*i.e.* not only the corresponding ones). Only a few segmentations had poor (<70%) to no overlap with the annotations: 3/47 for Seg1 and 2/42 for Seg2. The median semi-Dice, mean shortest distance and overlap grading (Table 1) further confirm the high accuracy of the segmentations on all regions. Figure 5 aimed to corroborate the relation between the visual gradings and the overlap measures. We refrained from statistically assessing the relation

Table 1. Median values and interquartile range of the semi-Dice coefficient, mean shortest distance and visual grade comparing all annotated regions Ano1 and Ano2 to the segmentations Seg1 and Seg2

Grading metric	Ano1-Seg1 ($n = 47$)	Ano2-Seg2 ($n = 42$)
Median semi-Dice	0.90 (0.82; 0.94)	0.84 (0.74; 0.90)
Median distance (mm)	1.45; (1.29; 1.63)	1.61 (1.36; 2.0)
Median grade (a.u.)	3 (3; 3)	3 (3; 3)

a.u., arbitrary units.

Table 2. Intraclass correlation coefficients of manual thickness measurements Obs1, Obs2 and Obs3 and semi-automatic measurements M1 and M2 for the corresponding regions ($n = 30$)

Observer	Obs1	Obs 2	Obs 3	M1	M2
Obs 1	1	0.342	0.516	0.542	0.453
Obs 2	0.342	1	0.603	0.617	0.529
Obs 3	0.516	0.603	1	0.737	0.738
M1	0.542	0.617	0.737	1	0.897
M2	0.453	0.529	0.738	0.897	1

because there are hardly any gradings in categories "0" and "1" and relatively few in category "2".

The overall variance of the manual thickness measurements was 2.9 mm^2; the variance of semi-automated measurement was significantly smaller: 0.46 mm^2 ($p = 1.1 \times 10^{-7}$). Most importantly, the ICC of the semi-automatic thickness measurements was found to be significantly higher than the ICC of the manual measurement ($p = 0.005$). This clearly demonstrates the limitation of performing manual thickness measurements and the benefit of semi-automatic measurement.

Previous studies have reported varying agreement in manual measurements of the bowel wall thickness.[7,11] Part of our data is also included in BLINDED.[7] In that study, the ICC was reported to be very good (0.87) for (two) experienced observers while it was reported to be good (0.69) for a mixed pool of (four) observers. On the other hand, a moderate ICC of 0.51 was reported in another study of 33 patients.[11] To our knowledge, all the previous studies evaluated the thickness measurement on all bowel segments of patients in a study population. As such, these measurements probably included a large number of healthy segments which could skew the ICC to higher values. Our thickness measurements were made on segments that were perceived as active by two experienced radiologists. In our opinion, the lower overall ICC than previously reported for BLINDED data (0.45 *vs* 0.69) shows the difficulty of objectively measuring the thickness of bowel segments with active disease on MRI.

Semi-automatic methods for segmentation have been widely employed for many other challenging problems in medical imaging such as lymph node detection,[23] segmentation of skin lesions,[24] tumour identification,[25] and organ localization and segmentation.[26]

Table 3. Intraclass correlation coefficients of manual thickness measurements Obs1, Obs2 and Obs3 and semi-automatic measurements M1 ($n = 47$) for all segments with an annotation Ano1

Observer	Obs 1	Obs2	Obs 3	M1
Obs 1	1	0.475	0.473	0.541
Obs 2	0.475	1	0.451	0.593
Obs 3	0.473	0.451	1	0.722
M1	0.541	0.593	0.722	1

Table 4. Intraclass correlation coefficients of manual thickness measurements Obs1, Obs2 and Obs3 and semi-automatic measurements M2 ($n = 42$) for all segments with an annotation Ano2

Observer	Obs 1	Obs2	Obs 3	M2
Obs 1	1	0.325	0.584	0.485
Obs 2	0.325	1	0.658	0.545
Obs 3	0.584	0.658	1	0.702
M2	0.485	5.45	0.702	1

These data illustrate the wide availability of techniques for (semi-) automatically segmenting abnormal regions.

There is almost no prior work on (semi-)automatic segmentation of the bowel wall, particularly in relation to Crohn's disease. In our opinion, this highlights the difficulty of the task.

Related in part, Bhushan et al[27] developed a motion correction and pharmacokinetic parameter estimation technique for identifying colorectal cancer using dynamic contrast-enhanced MRI data. Furthermore, Schunk et al[28] analyzed MR images for their suitability in analyzing inflammatory bowel diseases, including Crohn's disease. However, they did not explore computational tasks but instead focused on the clinical aspects.

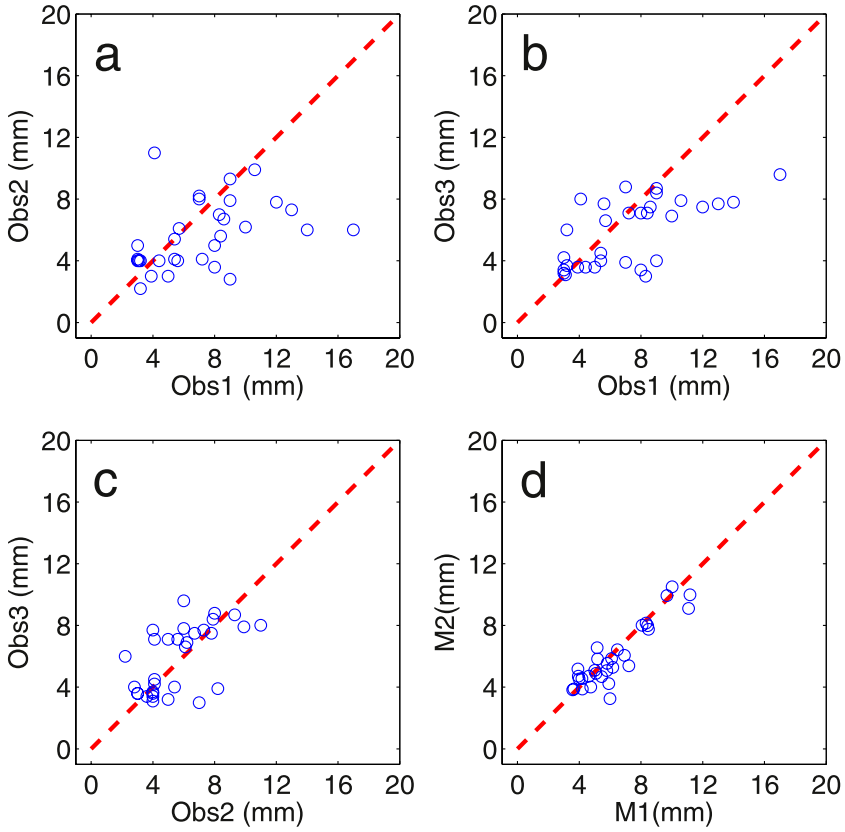
Very recently, Hampshire et al[18] introduced a sophisticated method for semi-automated measurement based on a technique called particle filtering. The article reports similar variability between the algorithm and observers in comparison to the variability between observers. Measurements were not particularly made in active segments but rather in random picked locations evenly distributed in the large bowel. Furthermore, the evaluation of the technique involved automated measurements that were closest to manual measurements and thus not independent. These differences likely enhance the agreement and prevent a comparison with our outcomes.

Lately, our group has developed a supervised learning framework for automatic detection and segmentation of Crohn's disease from abdominal MR images.[29] However, this method does not deliver a comprehensive segmentation of the colon wall but identifies only small regions sized several cubic millimetres affected by Crohn's disease.

The eventual thickness measurement was fully dependent on the preceding segmentation step. Additionally, future research might use segmentations to derive other features, *e.g.* the volume of the diseased region. For these reasons, we chose to separately evaluate the method regarding its performance (1) in segmenting a diseased part of the bowel wall and (2) in measuring a bowel segment's wall thickness.

Our work has several limitations. First, we focused on segments with active Crohns' disease on MRI only. We took this approach because these segments are most relevant both for diagnosis and assessment of disease severity. As a consequence, there were

Figure 6. Scatter plots depicting the manual and the semi-automatic thickness measurements for the 30 corresponding segments against each other.



segments presenting active disease on endoscopy that are not included in the measurements. Simultaneously, several of the measured segments with active endoscopic disease did have a normal appearance on the T_1 images. All these segments were graded with a "3" by our annotators (*i.e.* complete overlap of segmentation and annotation). As such, the segmentation of these normal-looking segments was also accurate. We expect that including normal segments will only positively skew the ICCs.

Second, we did not correlate the thickness measurements to endoscopic outcome measures. Conventionally, the thickness measurement is combined with other features in disease-grading systems which are validated against the colonoscopic assessment.[4,6,8–10] We consider such a validation outside the scope of our current article, but it is an important aspect in a comprehensive analysis of the prospective data that is in progress (publication in preparation).

In conclusion, our data show that a semi-automatic measurement technique facilitates a highly reproducible delineation of a region with active Crohn's disease. Furthermore, the semi-automatic thickness measurement achieves a significantly higher interobserver agreement than manual observers in active segments on MRI. As such, it may reduce the interobserver variability of MRI grading systems for Crohn's disease.

FUNDING

The research leading to these results was partly funded by the European Community's Seventh Framework Programme (FP7/2007–2013): the VIGOR++ Project (grant agreement no. 270379).

REFERENCES

1. Gourtsoyiannis NC, Grammatikakis J, Papamastorakis G, Koutroumbakis J, Prassopoulos P, Rousomoustakaki M. Imaging of small intestinal Crohn's disease: comparison between MR enteroclysis and conventional enteroclysis. *Eur Radiol* 2006; **16**: 1915–25. doi: https://doi.org/10.1007/s00330-006-0248-8

2. Tillack C, Seiderer J, Brand S, Goke B, Reiser MF, Schaefer C. Correlation of magnetic resonance enteroclysis (MRE) and wireless capsule endoscopy (CE) in the diagnosis of small bowel lesions in Crohn's disease. *Inflamm Bowel Dis* 2008; **14**: 1219–28. doi: https://doi.org/10.1002/ibd.20466

3. Albert JG, Martiny F, Krummenerl A, Stock K, Lesske J, Gobel CM. Diagnosis of small bowel Crohn's disease: a prospective comparison of capsule endoscopy with magnetic resonance imaging and fluoroscopic enteroclysis. *Gut* 2005; **54**: 1721–7. doi: https://doi.org/10.1136/gut.2005.069427

4. Steward MJ, Punwani S, Proctor I. 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. doi: https://doi.org/10.1016/j.ejrad.2011.07.013
5. Ziech ML, Bossuyt PM, Laghi A, Lauenstein TC, Taylor SA, Stoker J. Grading luminal Crohn's disease: which MRI features are considered as important? *Eur J Radiol* 2012; **81**: 467–72.
6. Rimola J, Rodriguez S, García-Bosch O, Ordás I, Ayala E, Aceituno M. Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. *Gut* 2009; **58**: 1113–20. doi: https://doi.org/10.1136/gut.2008.167957
7. Tielbeek JA, Makanyanga JC, Bipat S, et al. Grading Crohn Disease Activity With MRI: Interobserver Variability of MRI Features, MRI Scoring of Severity, and Correlation With Crohn Disease Endoscopic Index of Severity. *American Journal of Roentgenology.* 2013; **201**(6): 1220–1228.
8. Rimola J, Ordas I, Rodriguez S, Garcia-Bosch O, Aceituno M, Llach J. Magnetic resonance imaging for evaluation of Crohn's disease: validation of parameters of severity and quantitative index of activity. *Inflamm Bowel Dis* 2011; **37**: 1759–68. doi: https://doi.org/10.1002/ibd.21551
9. Buisson A, Joubert A, Montoriol PF, Da Ines D, Hordonneau C, Pereira B. Diffusion weighted magnetic resonance imaging for detecting and assessing ileal inflammation in Crohn's disease. *Aliment Pharmacol Ther* 2013; **37**: 537–45. doi: https://doi.org/10.1111/apt.12201
10. Hordonneau C, Buisson A, Scanzi J, Goutorbe F, Pereira B, Borderon C. Diffusion-weighted magnetic resonance imaging in ileocolonic Crohn's disease: validation of quantitative index of activity. *Am J Gastroenterol* 2013; **109**: 89–98. doi: https://doi.org/10.1038/ajg.2013.385
11. Siddiki HA, Fidler JL, Fletcher JG. Prospective comparison of state-of-the-art MR enterography and CT enterography in small-bowel Crohn's disease. *AJR Am J Roentgenol* 2009; **193**: 113–21. doi: https://doi.org/10.2214/ajr.08.2027
12. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. *Int J Computer Vis* 1988; **1**: 321–31. doi: https://doi.org/10.1007/bf00133570
13. Mumford D, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems. *Commun Pure Appl Mathematics* 1989; **42**: 577–658. doi: https://doi.org/10.1002/cpa.3160420503
14. Li C, Kao C, Gore JC, Ding Z. Implicit active contours driven by local binary fitting energy. Paper presented at: computer vision and pattern recognition CVPR 2007. doi: https://doi.org/10.1109/cvpr.2007.383014
15. Wang L, He L, Mishra A, Li C. Active contours driven by local Gaussian distribution fitting energy. *Signal Process* 2009; **86**: 2345–447.
16. Lynch M., Ghita O, Whelan PF. Segmentation of the left ventricle of the heart in 3-D+t MRI data using an optimized nonrigid temporal model. *IEEE Trans Med Imaging* 2008; **27**: 195–203. doi: https://doi.org/10.1109/tmi.2007.904681
17. Wang L, Shi F, Weili L, Gilmore JH, Shen D. Automatic segmentation of neonatal images using convex optimization and coupled level sets. *Neuroimage* 2011; **58**: 805–17. doi: https://doi.org/10.1016/j.neuroimage.2011.06.064
18. Hampshire T, Menys A, Jaffer A. A probabilistic method for estimation of bowel wall thickness in MR colonography. *PLoS One* 2017; **12**: e0168317.
19. Ziech ML, Lavini C, Caan MW, et al. Dynamic contrast-enhanced MRI in patients with luminal Crohn's disease. *Eur J Radiol.* 2012; **81**(11): 3019–3027.
20. Ziech ML, Stoker J. MRI of the small bowel: enterography. *MRI of the gastrointestinal tract*. Berlin, Germany: Springer-Verlag; 2010.
21. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; **26**: 297–302. doi: https://doi.org/10.2307/1932409
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74.
23. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D. Automatic detection and segmentation of lymph nodes for ct data. *IEEE Trans Med Imaging* 2012; **31**: 240–50. doi: https://doi.org/10.1109/tmi.2011.2168234
24. Pereyra N, Dobigeon N, Batatia H, Tourneret JY. Segmentation of skin lesions in 2-d and 3-d ultrasound images using spatially coherent generalized Rayleigh mixture model. *IEEE Trans Med Imaging* 2012; **31**: 1509–20. doi: https://doi.org/10.1109/tmi.2012.2190617
25. Song Y, Cai W, Kim J, Feng D. A multistage discriminative model for tumor and lymph node detection in thoracic images. *IEEE Trans Med Imaging* 2012; **31**: 1061–75.
26. Criminisi A, Robertson D, Konukgolu E. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Med Image Anal* 2013; **17**: 1293–303. doi: https://doi.org/10.1016/j.media.2013.01.001
27. Bushan M, Schnabel J, Risser L, Heinrich M, Brady JM, Jenkinson M. Motion correction and parameter estimation in DCE-MRI sequences: application to colorectal cancer. *Med Image Comput Comput Assist Interv* 2011; **14**: 476–83.
28. Schunk K. Small bowel magnetic resonance imaging for inflammatory bowel disease. *Top Magn Reson Imaging* 2002; **13**: 409–25. doi: https://doi.org/10.1097/00002142-200212000-00005
29. Mahapatra D, Schuffler PJ, Tielbeek JAW, et al. Automatic Detection and Segmentation of Crohn's Disease Tissues from Abdominal MRI. *IEEE Transactions on Medical Imaging.* 2013; **32**(12): 2332–2348.
30. Caselles V, Kimmel R, Sapiro G. Geodesic active contours. *Int J Computer Vis* 1997; **22**: 61–79.

APPENDIX A

SCAN PROTOCOL

Table A1. Scan protocol applied to acquire the prospective data

	Plane	Matrix	Slice thickness	FOV (mm^2)	TR (ms)	TE (ms)	Flip angle (ms)
Balanced GE	Coronal	400 × 400	5	380 × 380	2.5	1.25	60
BTFE dynamic fat sat	Coronal	192 × 192	10	380 × 380	2.0	1	45
T_2-SSFSE	Coronal	400 × 400	4	380 × 380	6660	60	90
T_2-SSFSE	Axial	528 × 528	4	400 × 400	759	119	90
T2-w SSFSE fat sat	Axial	320 × 320	7	380 × 302	1314	50	90
DCE sequence	Coronal	192 × 224 × 30	NA	380 × 439	2.9	1.8	15
3D T1-w SPGE fat sat	Coronal	200 × 240 × 90	NA	380 × 459	2.2	1.0	10
3D T1-w SPGE fat sat	Axial	384 × 384 × 90	NA	380 × 380	2.1	1.0	10

3D, three-dimensional; BTFE, balanced turbo field-echo; DCE, dynamic contrast enhanced; fat sat, fat saturated; FOV, field of view; GE, gradient echo; NA, not applicable; SPGE, spoiled gradient-echo; SSFSE, single-shot fast spin echo; T1-w, T_1 weighted; T2-w, T_2 weighted; TE, echo time; TR, repetition time.

APPENDIX B

TECHNICAL DESCRIPTION OF THE SEGMENTATION ALGORITHM

The bowel wall's inner surface was segmented using an active contours algorithm, which extends the well-known local binary fitting approach.[14,15] This algorithm was initialized by computing the signed distance transform to the boundary obtained by dilating a manually drawn centreline using a small structuring element of $3 \times 3 \times 3$ voxels. A negative distance to the contour denoted the inside and a positive distance denoted the outside of the lumen. This level-set function evolved to optimize an energy function E_{totI}, which balanced a data term E_{dataI} and a shape term E_{shapeI}:

$$E_{\text{totI}} = E_{\text{dataI}} + E_{\text{shapeI}} \tag{A1}$$

The data term took into account that the lumen as well as the exterior could contain materials with varying intensity. Therefore, the intensity $I(y)$ in a neighbourhood around x was modelled by a locally varying function $\xi_i(x,y)$:

$$E_{\text{dataI}} = \int_{\Omega} \sum_{i=\{\text{lumen,background}\}} \int_{\Omega} K_{\rho}(x-y), \left(\frac{[I(y) - \xi_i(x,y)]^2}{2\eta_i^2(x)} + \log \eta_i(x) \right) M_i[\phi_A(y)] \mathrm{d}y \mathrm{d}x \tag{A2}$$

In this equation, the outer integral sums over the entire image domain $\Omega \subset R^3$, so that x represents an image co-ordinate. The summation is over the lumen and the background separately, while an indicator function $[M_i(z)]$ is applied that is one if $z \in \Omega_i$ and is zero if otherwise. As such, the inner integral represents a convolution sum that only takes into account terms from the region indexed by the indicator function. Furthermore, $x - y$ is a neighbourhood co-ordinate and $K_{\rho}(x-y)$ a weight function (*e.g.* a Gaussian) imposing locality around voxel x. Finally, $\eta_i^2(x)$ is the residual with regard to the model (*i.e.* the local variance).

Essentially, $\xi_i(x,y)$ can take on an intensity that depends on the material encountered in y. This can correspond to water, air or faecal material for the lumen. Similarly, dark, intermediate and bright intensities are assumed to occur in the background. As such, we were able to cope with the varying constituency on either side of the level-set function.

The shape term served for regularization and consisted of the commonly used weighted minimal length term:[30]

$$E_{\text{shapeI}} = \int_{\Omega} g(x, \phi_A) |\nabla \phi(x)| \delta_{\in}[\phi_A(x)] \mathrm{d}x \tag{A3}$$

where $\delta_{\in}$ represents a regularized Dirac function and $g(x)$ is a weight function that is low whenever the gradient of the level-set function is directed to the normal of the image gradient.

The total energy function was optimized iteratively by a two-step process: (1) $\xi_i(x,y)$ and $\eta_i(x)$ were updated by minimizing Equation (A2) for a fixed ϕ_A; and (2) $\phi_A(x)$ was adjusted using a gradient descent approach.

This optimization resulted in a segmentation of the inner surface of the bowel wall, which was fixated. Subsequently, a similar active contour approach was taken to coarsely segment the outer surface of the bowel wall with a second level-set function. This second level-set function was initialized by outwardly dilating the final segmentation of the inner surface by 4 mm, approximately the average thickness of

Table A2. Scan protocol applied to acquire the retrospective data

	Plane	Matrix	Slice thickness (mm)	FOV	TR (ms)	TE (ms)	Flip angle
T_2-SSFSE	Coronal	256 × 256	4	400 × 400	516–758	65–118	90
T_2-SSFSE	Axial	256 × 256	4	400 × 400	516–758	65–118	90
T2-w SSFSE fat sat	Axial	288 × 288	7	375 × 300	1370–1450	70	90
DCE sequence	Coronal	144 × 144 × 14	NA	400 × 400 × 35	2.9	1.8	6
3D T1-w SPGE fat sat	Coronal	192 × 192 × 100	NA	400 × 400 × 200	1.87–2.19	1.0	10
3D T1-w SPGE fat sat	Axial	208 × 208 × 70	NA	400 × 400 × 140	1.87–2.19	1.0	10

DCE, dynamic contrast enhanced; fat sat, fat saturated; FOV, field of view; NA, not applicable; SPGE, spoiled gradient-echo; SSFSE, single-shot fast spin echo; T1-w, T_1 weighted; T2-w, T_2 weighted; TE, echo time; TR, repetition time.

the healthy bowel wall.[20] The ensuing boundary again served to generate a level-set function $\phi_B(x)$ in which a positive distance to the contour denoted the inside (wall) and a negative distance denoted the background. The level-set again evolved to optimize an energy function: $E_{\text{totO}} = E_{\text{dataO}} + E_{\text{shapeO}}$.

The data term of the second energy function was very similar to the one segmenting the bowel lumen except for that the bowel wall was assumed to contain only one component: $\xi_{\text{wall}}(x,y) = \mu_{\text{wall}}(x)$. Therefore, the indicator function $[M_{\text{wall}}(z)]$ was adjusted such that it only took into account terms corresponding to the wall (and discarded terms if z was in the lumen).

The shape term E_{shapeO} consisted of three terms:

$$E_{\text{shapeO}} = \lambda_1 E_{\text{regularO}} + \lambda_2 E_{\text{thickO}} + \lambda_3 E_{\text{crossO}} \qquad \text{(A4)}$$

The first term in Equation (A4) was the same minimal length term as applied to the inner surface segmentation $[E_{\text{regularO}} = E_{\text{shapeI}}(\phi_B)]$. The second term targeted to keep the outer surface close to the average wall thickness:

$$E_{\text{thickO}} = \int_{\Omega} \left[\phi_A(x) \delta_{\in}(\phi_B)(x) - \mu_{\text{average}} \right]^2 \mathrm{d}x \qquad \text{(A5)}$$

which returns the squared deviation from the average bowel wall thickness integrated along the outer surface.

The third term of Equation (A4) merely prevented the inner and the outer surface representations from crossing each other:

$$E_{\text{cross}}(\phi_A, \phi_B) = \int_{\Omega} H(\phi_A(x)) H(\phi_B(x)) \mathrm{d}x \qquad \text{(A6)}$$

in which H represents a heavy side function. Essentially, Equation (A6) returns a (positive) number if the outer surface crosses the inner surface and is zero if otherwise. Setting λ_3 to a very large value precludes such a crossing.

The energy function for the bowel wall's outer surface segmentation was optimized following the same strategy as was followed for the inner surface segmentation. The two resulting level-sets together yielded an accurate delineation of the bowel wall.