Assessing Small Bowel Stricturing and Morphology in Crohn's Disease Using Semi-automated Image Analysis

Ryan W. Stidham, MD, MS,*,†,‡ Binu Enchakalody, MS,†,¶ Akbar K. Waljee, MD, MSc,*,‡,§ Peter D. R. Higgins, MD, PhD, MSc,* Stewart C. Wang, MD, PhD,†,¶ Grace L. Su, MD,*,†,§ Ashish P. Wasnik, MD,†,‖ and Mahmoud Al-Hawary, MD*,†,‖

Background: Evaluating structural damage using imaging is essential for the evaluation of small intestinal Crohn's disease (CD), but it is limited by potential interobserver variation. We compared the agreement of enterography-based bowel damage measurements collected by experienced radiologists and a semi-automated image analysis system.

Methods: Patients with small bowel CD undergoing a CT-enterography (CTE) between 2011 and 2017 in a tertiary care setting were retrospectively reviewed. CT-enterography studies were reviewed by 2 experienced radiologists and separately underwent automated computer image analysis using bowel measurement software. Measurements included maximum bowel wall thickness (BWT-max), maximum bowel dilation (DIL-max), minimum lumen diameter (LUM-min), and the presence of a stricture. Measurement correlation coefficients and paired t tests were used to compare individual operator measurements. Multivariate regression was used to model identification of strictures using semi-automated measures.

Results: In 138 studies, the correlation between radiologists and semi-automated measures were similar for BWT-max ($r = 0.724, 0.702$), DIL-max ($r = 0.812, 0.748$), and LUM-min ($r = 0.428, 0.381$), respectively. Mean absolute measurement difference between semi-automated and radiologist measures were no different from the mean difference between paired radiologists for BWT-max (1.26 mm vs 1.12 mm, $P = 0.857$), DIL-max (2.78 mm vs 2.67 mm, $P = 0.557$), and LUM-min (0.54 mm vs 0.41 mm, $P = 0.596$). Finally, models of radiologist-defined intestinal strictures using automatically acquired measurements had an accuracy of 87.6%.

Conclusion: Structural bowel damage measurements collected by semi-automated approaches are comparable to those of experienced radiologists. Radiomic measures of CD will become an important new data source powering clinical decision-making, patient-phenotyping, and assisting radiologists in reporting objective measures of disease status.

Key Words: radiomics, computer-aided image analysis, segmentation, intestinal fibrosis, CT enterography, MR enterography

INTRODUCTION

Image-based disease assessments provide essential diagnostic and prognostic information for patients with small bowel Crohn's disease (CD).[1] Presently, therapeutic trials and proactive clinical management of CD focuses on the inflammatory measures of disease activity such as mucosal healing, C-reactive protein, and fecal calprotectin. However, pre-existing structural bowel damage can be a major limitation of anti-inflammatory therapy and is a common reason for medical failure despite apparent anti-inflammatory efficacy.[2, 3] Cross-sectional imaging using CT-enterography (CTE) or MR-enterography (MRE) can assess deep bowel structural damage and has been shown to predict long-term outcomes in CD.[4, 5] Further, there is evidence supporting substantial additional diagnostic yield of imaging compared with clinical or endoscopic assessment alone based on visualization of the bowel wall and surrounding tissues.[6, 7] As a result, imaging features are being investigated as therapeutic endpoints in preparation for future studies of antifibrotic medications.[8, 9]

Similar to endoscopic assessments, information gleaned from IBD imaging needs to overcome issues of subjectivity, interobserver variation, and nonstandardized reporting. Recognizing these challenges, the American Gastrointestinal Association and Society of Abdominal Radiology joint society released a consensus statement on IBD imaging and reporting that aims to clarify validated imaging feature definitions and criteria for disease activity reporting based on a morphologic construct.[10] Additionally, central reading of enterography studies has demonstrated agreement between radiologists that is very good for overall disease activity scoring (intraclass correlation [ICC] = 0.84), good for quantitative bowel wall thickness (ICC = 0.74), and fair for contrast enhancement measurements (ICC = 0.59).[11] Accurate and reproducible measurement of disease characteristics are possible, yet interpretation still relies on highly trained expert radiologists who are in limited supply. Detailed imaging assessment for every IBD enterography performed is achievable with central reading in clinical trials but may be less feasible in clinical practice considering the time needed for additional measurements, local availability of IBD imaging expertise, and challenges of standardizing measurement protocols.

Advances in image analysis techniques offer the opportunity to improve measurement quality outside the confines of resource-rich clinical trials. Automated machine-based assessments of small bowel morphology could be valuable for measurement standardization, reproducibility, and making the availability of measurements more feasible in community practice settings. Further, subjectively identified and graded features, such as the presence of strictures, could be standardized using data modeling. Methods that quantitatively capture image-based descriptions of disease—the burgeoning field of Radiomics—offer an expansive information source for population-based big-data analyses. Here, we compare the agreement of small bowel structural damage measurements on CT-enterography studies between semi-automated image analysis techniques and 2 abdominal radiologists with expertise in IBD.

METHODS

Subject Selection

In this HIPAA-compliant, retrospective study approved by the University of Michigan' Institutional Review Board, patients with CD who underwent a CTE between 2011 and 2017 were identified from the electronic health records at a tertiary care center in the United States. A clinical diagnosis of CD was determined using a previously validated definition requiring 2 ICD-9 or ICD-10 (after October 1, 2015) diagnosis codes for CD on 2 separate encounters and at least 1 record of CD medication use.[12] Because of our focus on evaluation of small bowel CD, included subjects were required to have a history of

>5 cm of small bowel disease within the ileum based on review of radiographic reports. Studies were excluded for radiographic evidence of colonic disease, active penetrating disease (enteral abscess or fistula), an ileostomy, or prior surgery. If multiple CTE studies were available for an individual subject, the first CTE performed was selected for analysis to avoid correlated studies. Subject demographics, medication use, and laboratory values at the time of scan were extracted by electronic medical record review.

CD Structural Feature Reference Measurements

Studies underwent manual measurement of structural intestinal characteristics by 2 fellowship-trained abdominal radiologists with expertise in inflammatory bowel disease and over 10 years' experience. Radiologists were blinded to subject identifiers, clinical information, original radiologist interpretation, and each other's measurements. Original CTE Digital Imaging and Communications in Medicine (DICOM) files were reviewed in traditional axial, coronal, and sagittal projections with 1 to 2 mm reconstructed slice thickness. A custom graphical user interface (GUI) was used to collect manual radiographic measurements using electronic calipers.

Select structural features were measured including maximum bowel wall thickness (BWT-max), maximum bowel dilation diameter (DIL-max), and minimal lumen diameter (LUM-min) within the small intestine proximal to the ileocolonic junction. Measurements were collected using the custom GUI, allowing recording of both traditional linear measurements and the location selected for measurement. Direct maximum bowel wall thickness measurements were also classified as mild (3 to 5 mm), moderate (5 to 9 mm), or severe (>9 mm) based on recent guidance from the American Gastrointestinal Association (AGA) and Society of Abdominal Radiology (SAR).[10]

Study radiologists were also asked to subjectively identify the presence, suspicion, or absence of small bowel strictures with and without upstream small bowel dilation. A recent expert consensus report led by Rieder et al suggested the following criteria for defining intestinal strictures: localized luminal narrowing and bowel wall thickening, at least 25% increase in thickness of maximally thickened bowel, luminal diameter reduction of at least 50% relative to adjacent normal bowel loop, and upstream small bowel dilation of 30 mm or more.[9]

Semi-automated Bowel Morphomics Measurements

The automated measurements of small intestinal structural features relied on bowel segmentation. Both study radiologists independently placed a series of reference points for each study, beginning at the ileocolonic junction and covering the entire diseased region of bowel, terminating at approximately

10 cm proximal to end of the diseased segment using a custom-designed DICOM viewer tool. The interpolated centerline was then processed using curved planar reformation (CPR) with the centerline as the center of the volume to generate a tube-like straightened reconstruction of the originally convoluted bowel segment (Fig. 1). Next, bowel outer and inner wall segmentation was performed by iteratively modeling a polynomial grid on the transition edges between wall to mesentery and lumen to wall using super-pixel voxel segmentation followed by k-means classification. The best-fit grids for the outer and inner wall were remapped to the original DICOM image on cartesian coordinates and saved as 3D contours. The resulting segmented intestine underwent quality control review for segmentation errors.

Using the bowel wall and lumen masks, we measured the bowel thickness volume by excluding the lumen volume from the total bowel volume. The total bowel wall thickness and bowel lumen area profiles were used to derive descriptive statistics including maximum, minimum, median, and mean radius and diameter for each feature continuously along the length of the segmented intestine (Fig. 2). Because of the irregular shape of intestine (eg, an imperfect ellipse), semi-automated measures report the equivalent radius or diameter of the perfect circle generated using the total cross-sectional area of lumen, bowel wall, or entire intestinal cross-section.

Data Analysis

Descriptive statistics were provided for both radiologist-acquired and semi-automated bowel measurements. Measurements were compared between radiologists to establish paired reviewer agreement for each feature. Because of imperfect reference radiologist agreement, semi-automated measures were compared with both the mean of radiologists' measurements and with those of the individual radiologists. The Pearson correlation coefficient was used to compare the paired continuous measurements. The paired sample t test was used to compare measurement differentials between radiologists and semi-automated measures to detect statistically significant positive or negative measurement bias for any study reviewer. Interclass correlation coefficient (ICC) was used to assess the agreement between automated measures derived from different radiologist reference point placements. Multivariate logistic regression using only automated measurements was used to model the presence of an intestinal stricture based on radiologist assessment. Models used automatic backwards variable selection with no forced variables. Intestinal stricture detection model area under the receiver operating curve (AuROC), accuracy, sensitivity, and specificity were reported. Statistical analysis was performed using SAS 9.4 (SAS; Cary, NC).

RESULTS

Patient Characteristics

A total of 138 studies met selection criteria and underwent radiologist review and image analysis; patient characteristics at the time of imaging are shown in Table 1. The average patient age was 43.9 years, and females comprised 51.4% of the cohort. Nearly half of participants were using immunomodulators, whereas only 25% were using a biologic
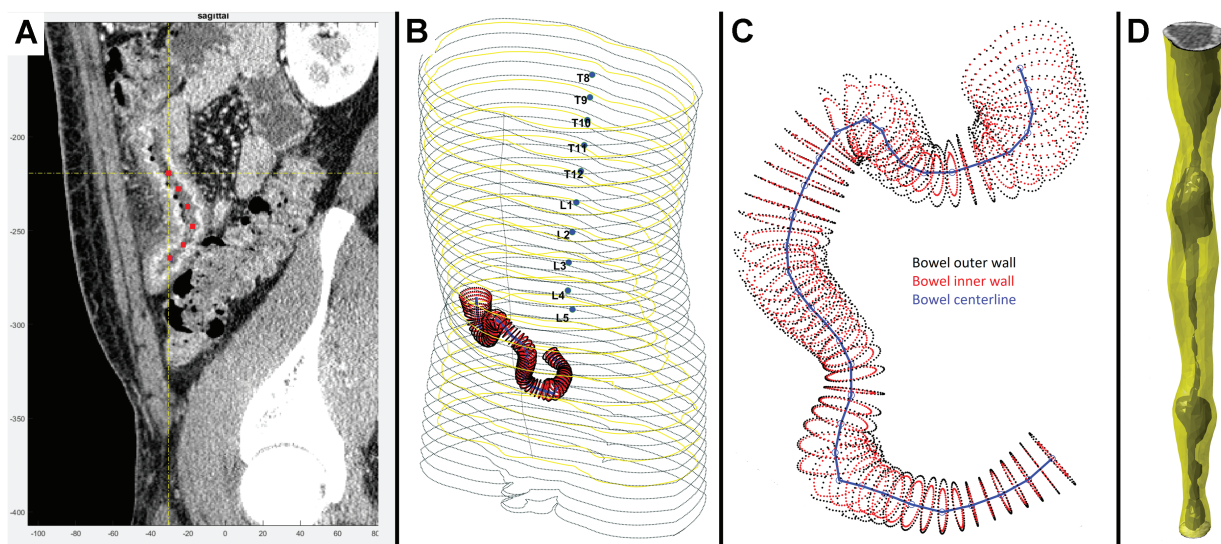
FIGURE 1. Illustration of bowel segmentation and reconstruction process. A, Radiologist places reference center points to identify distal small bowel. B, Centerline within small bowel is shown in 3D space in the context of patient abdominal silhouette. C, Centerline shown with automatically identified inner and outer bowel wall perimeters using superpixel voxel segmentation followed by k-means classification, allowing segmentation of lumen, bowel wall, and extra-intestinal space. D, Example of straightened tube-like reconstruction of segmented intestine with bowel wall and shaded lumen identified.
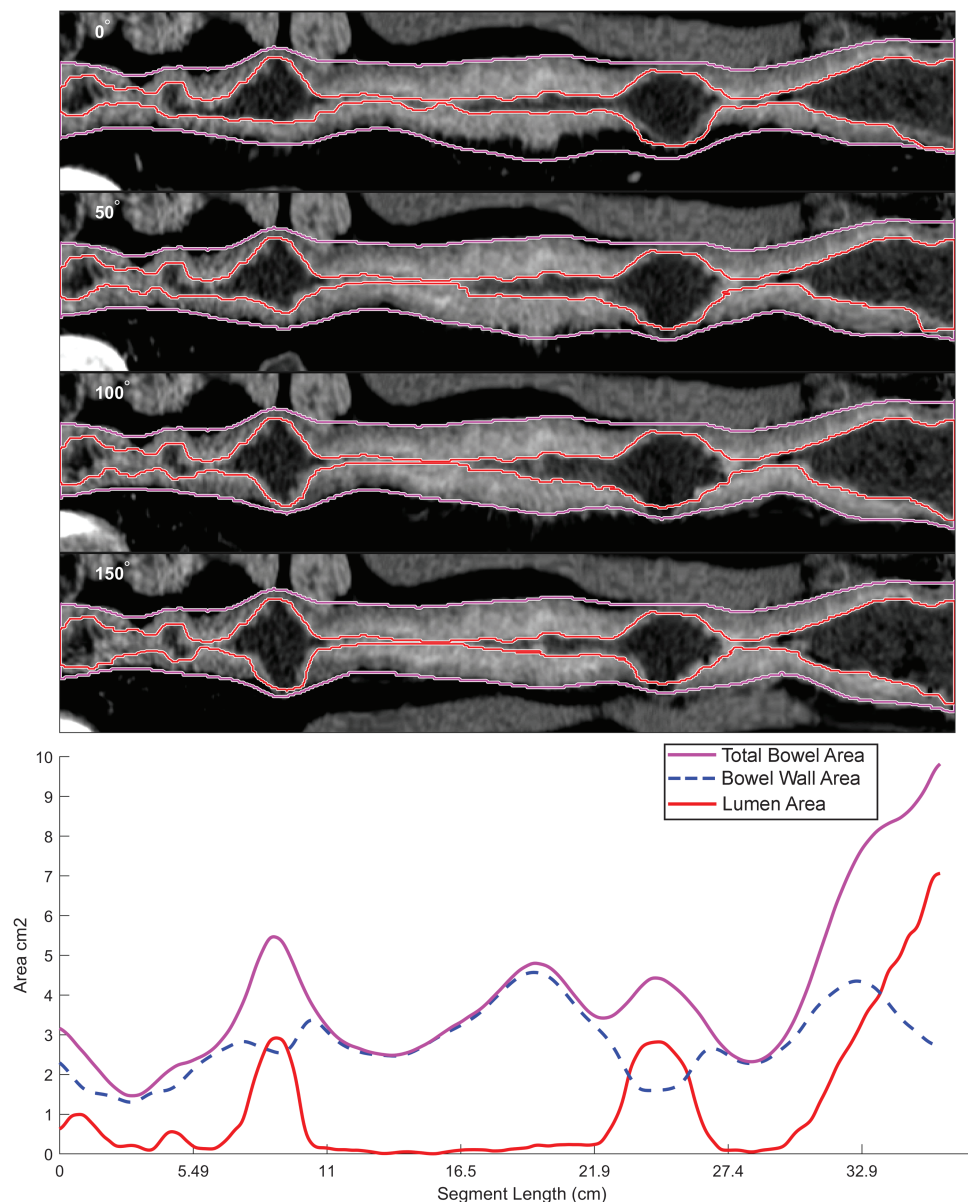
FIGURE 2. Example of small bowel segmentation and quantitative output. A representation of small bowel segmentation, with total bowel cross-sectional area, bowel wall area, and lumen area, expressed as centimeters squared. Projections of segmented bowel rotated about the lumen centerline at 0, 50, 100, and 150 degrees are shown; data from all 360-degree rotational projections are incorporated into bowel measurements. Data can also be exported into tabular format. The conversion from cross sectional area into an ideal equivalent diameter (for bowel dilation and lumen measurements) is the square root of cross-sectional area divided by π.

therapy at the time of imaging. The mean C-reactive protein (CRP) within 4 weeks of imaging was elevated at 1.6 mg/dL (normal <0.6 mg/mL) with a maximum of 10.3 mg/mL. When examining the mean radiographic features of the cohort as measured by the study radiologists, average BWT-max was 9.4 mm (SD 2.6 mm), and average lumen minimum diameter was 2.7 mm (SD 1.2 mm). Mean small bowel maximum dilation diameter was 23.8 mm (SD 6.8 mm), with minimum and maximum values of 11.0 mm and 44.3 mm, respectively. Regarding

suggested thresholds for clinically significant upstream bowel dilation, subjects demonstrating >25 mm, >30 mm, or >35 mm maximum diameter comprised 41.3%, 21.0%, and 13.1% of the cohort, respectively. The cohort was skewed toward moderate and severe AGA-SAR BWT severity classes, with only 6.4% of the cohort classified as mild (3 to 5 mm), 51.2% moderate (5 to 9 mm), and 42.4% severe (>9 mm). The median length of segmented bowel for automated analysis was 28.4 cm (range: 12.6 to 72.1 cm).

Maximum Bowel Wall Thickness Assessment

Maximum bowel wall thickness measurement correlation between radiologists was very good ($r = 0.724$, $P < 0.0001$). The mean absolute difference between radiologists' BWT-max measurements was 1.1 mm (SD 1.3 mm); no BWT measurement bias for either radiologist was detected ($P = 0.375$; Table 2). The correlation between semi-automated and mean radiologist BWT-max measurements was similar to that between 2 radiologists ($r = 0.702$ vs 0.724, respectively). Mean absolute difference between semi-automated and radiologist BWT-max measurement was 1.26 mm (SD 1.95 mm) without any detected measurement bias (0.07 mm; 95% CL [confidence limits], −0.31 mm to 0.37 mm; $P = 0.857$; Table 2). When classifying agreement, semi-automated and radiologist BWT measures were within 3, 2, or 1 mm in 89.6%, 79.2%, and 38.4% of cases, respectively, which was similar to the agreement between radiologists of 88.9%, 72.8%, and 44.8% (Table 3). Considering the challenge of establishing ground truth in the absence of exact agreement between reference radiologists, agreement between semi-automated measures and *either* radiologist A or B was explored. Semi-automated methods were within 3, 2, or 1 mm of at least 1 reference radiologist BWT measurement in 94.9%, 90.6%, and 58.7% of cases, respectively. Comparing automated BWT measures generated using reference points between the 2 study radiologists, measurement agreement was very good, with an ICC = 0.849.

Maximum Small Bowel Dilation Diameter

The mean absolute difference between radiologist maximum small bowel dilation measures was 2.7 mm (SD 2.0 mm), but a measurement bias of −0.63 mm (95% CL, −1.24 to −0.015 mm; $P = 0.045$) was present, indicating radiologist A made systematically smaller measurements compared with radiologist B. Agreement between radiologists for classifying the presence of small bowel dilation greater than 30 mm was good ($\kappa = 0.687$; 95% CL, 0.538 to 0.837), correctly classifying 75% of studies with that degree of dilation. Similar to BWT-max, semi-automated and mean radiologist DIL-max correlation ($r = 0.748$, $P < 0.0001$) was comparable with the correlation between radiologists ($r = 0.812$, $P < 0.0001$). Although semi-automated DIL-max exhibited a nonsignificant trend toward being larger than radiologists' measurements (+0.53 mm; 95% CL, −1.13 to 0.08; $P = 0.087$), that difference was similar to the differential between radiologists ($P = 0.557$, Table 2). Semi-automated DIL-max measures were within 5, 3, or 1 mm of mean reference radiologists in 72.7%, 52.5%, and 22.3% of cases (Table 3). When compared with either radiologist, semi-automated measurements were within 5,

TABLE 1. Patient Characteristics

	Mean	(SD), %
Age, yrs	43.9	16.2
Female (%)	71	51.4
Tobacco Use History	38	27.5
Medication Use		
Immunomod Use at Index Imaging (%)	72	52.2
Biologic Use at Index Imaging (%)	31	22.5
Labs		
White Blood Cell Count, kCell/mL	8.7	2.9
Hemoglobin, g/dL	13.4	2.0
Platelets, k/mL	306.2	91.8
C-Reactive Protein, mg/dL	1.6	2.3
Albumin, g/L	4.2	0.5
Radiologist Mean Measures		
Bowel Wall Thickness Maximum, mm	9.4	2.6
Bowel Dilation Maximum, mm	23.8	6.8
Lumen Minimum, mm	2.7	1.2

TABLE 2. Absolute Differences Between Radiologists and Semi-automated Measurements

[a]Radiologist vs Radiologist Measures

Measure	Mean Absolute Difference, mm	SD, mm	Minimum Difference, mm	Maximum Difference, mm	*P*
BWT-max	1.12	1.34	0.10	4.60	0.375
DIL-max	2.67	1.98	0.10	6.98	0.045
LUM-min	0.41	1.80	0.01	5.64	0.204

[b]Semi-automated vs Mean Radiologist Measures

Measure	Mean Absolute Difference, mm	SD, mm	Minimum Difference, mm	Maximum Difference, mm	*P*
BWT-max	1.26	1.95	0.20	6.70	0.857
DIL-max	2.78	1.81	0.07	6.90	0.557
LUM-min	0.54	1.90	0.02	6.14	0.596

[a]Absolute differences between paired radiologist measures are shown. *P* values listed reference the comparison of mean difference between radiologists to identify potential systematic measurement bias.

[b]Differences between semi-automated measurements and mean radiologist measurements are shown. *P* values shown reference the paired comparison of the difference between automated and mean radiologist measurement versus the measurement difference between radiologists for the same imaging study.

3, or 1 mm of either radiologist in 79.1%, 56.1%, and 27.3% of studies, respectively. Comparing automated DIL measures generated using reference points between the 2 study radiologists, measurement agreement was excellent, with an ICC = 0.900.

Minimum Lumen Diameter

Lumen minimum diameter (LUM-min) measurement correlation between radiologists was only fair ($r = 0.428$, $P < 0.0001$). The mean difference between radiologists LUM-min measurements was 0.41 mm (SD 1.8 mm), with no bias detected ($P = 0.204$). Semi-automated LUM-min measurements did exhibit a nonsignificant trend of being +0.40 mm (95% CL, −0.03 to 0.62; $P = 0.089$) greater than radiologist mean measurements. However, the correlation between semi-automated and mean radiologist LUM-min ($r = 0.381$, $P < 0.0001$) approximated the correlation between radiologists. Further, there was no difference in the disagreement between semi-automated and radiologist compared with paired radiologists measurements ($P = 0.596$). The LUM-min agreement within 3, 2, and 1 mm between paired radiologists compared with semi-automated vs radiologist measurements was similar (Table 3). When examining the difference of semi-automated measures to either radiologist's LUM-min, differences within 3, 2, or 1 mm improved to 78.4%, 61.2%, 30.2% of cases, respectively. Comparing automated LUM measures generated using reference points between the 2 study radiologists, measurement agreement was very good, with an ICC = 0.823.

Models for Identifying Subjectively Defined Stricturing Disease

Despite attempts to standardize definitions, discussion continues regarding the relative weighting of features that constitute the presence of an intestinal stricture. Applying suggested stricture definition criteria of at least a 25% increase in bowel wall thickness and a 50% reduction in lumen diameter relative to an adjacent normal small bowel loop, with 30 mm or more of upstream small bowel dilation, 15.2% had a stenosis identified by study radiologists. When using automatically acquired quantitative bowel measurements, modeling the presence of a radiologist-defined stricture being present had an AuROC of 0.857, with an accuracy, sensitivity, and specificity of 87.6%, 67.2%, and 92.5%, respectively; the model was unsurprisingly driven by maximum bowel dilation diameter (Table 4). Alternatively, when exploring identification of a probable stricture without the criteria of upstream dilation given the potential for decompression due to nasogastric tube placement, vomiting, or reduced oral intake, radiologists reported a suspected stricture in 46.7% of studies. The model of radiologist indication of a probable stricture being present demonstrated an AuROC of 0.917, with an accuracy, sensitivity, and specificity, of 84.4%, 95.6%, and 65.9%, respectively; the principal model component was minimal lumen diameter (Fig 4.).

DISCUSSION

Measurements describing structural bowel damage in Crohn's disease can be obtained from enterography studies using semi-automated image analysis methods. Automated measurements including bowel wall thickness, maximum bowel dilation diameter, and lumen diameter have similar correlation to reference radiologists compared with the agreement between 2 experienced radiologists. Although we expected the agreement of measurement values and colocalization to be related (between

TABLE 3. Classification of Agreement Between Reviewers and Automated Measures

Bowel Dilation Maximum	≤5 mm	≤3 mm	≤1 mm
Radiologist-Radiologist	76.9%	57.2%	20.7%
Radiologist-SemiAutoMeasures	72.7%	52.2%	22.3%
Bowel Wall Thickness Maximum	≤3 mm	≤2 mm	≤1 mm
Radiologist to Radiologist	88.9%	72.8%	44.8%
SemiAutoMeasures to Radiologist	89.6%	79.2%	38.4%
Lumen Diameter Minimum	≤3 mm	≤2 mm	≤1 mm
Radiologist-Radiologist	69.5%	41.2%	22.8%
Radiologist-SemiAutoMeasures	76.3%	52.3%	22.3%

Measurement differences were classified into clinically relevant groups to highlight degree of agreement between reference radiologists and semi-automated measures to mean radiologist measurements. Percentages shown represent the portion of all cases where paired measurements were within the indicated difference class.

TABLE 4. Modeling Radiologist Decision on Presence of an Intestinal Stricture Using Automated Measures

[a]Stricture Present, Defined as Relative Lumen Narrowing, Bowel Wall Thickening and Upstream Dilation

	Odds Ratio	95% Confidence Limits		P
DIL Max	1.22	1.10	1.36	<0.001
LUM Min	0.60	0.00	1.12	0.060

[b]Stricture Suspected, Upstream Bowel Dilation Not Required

	Odds Ratio	95% Confidence Limits		P
LUM min	0.37	0.02,	0.81	0.009
DIL max	2.60	1.51,	4.46	<0.001
BWT max	1.58	1.03,	2.41	0.034

[a]Models of radiologist identification of small bowel stricture using automatic bowel measurements had an AuROC of 0.857 with an accuracy of 87.6%. Bowel wall thickness was automatically dropped from this model due to co-linearity and no contribution to model fit as a result of maximum bowel diameter dominating model performance.

[b]Models of radiologists' identification of a possible stricture suspected, where explicit need for small bowel dilation >30 mm was not required, had an AuROC of 0.917 and accuracy of 84.4%.

both radiologists and automated image analysis methods), we found no correlation of spatial proximity and value for any of the structural features studied. Finally, we show that the identification of intestinal strictures—a radiographic finding that is often qualitatively defined—can be accurately modeled using automatically acquired quantitative image data.

These results add to the growing literature describing the use of computer-assisted image analysis to improve the reproducibility, objectivity, and potential accuracy of measurements where precision is valuable but tedious or, in some cases, not feasible to collect.[13–15] Crohn's disease structural damage scores, including the pioneering Lémann index, are dependent upon the objectivity and reproducibility of endoscopic and imaging assessments.[16] Prior work has demonstrated the potential for

inconsistencies of radiographic measurements, mainly bowel wall thickness, among less experienced providers.[17] However, BWT measurement agreement among highly experienced reviewers has been reported as having an ICC = 0.73, notably similar to agreement between radiologists and semi-automated methods in this study.[17, 18] These points highlight the benefits of central image review by unbiased trained experts, as recently described by Jairath et al in their work evaluating the feasibility of imaging endpoints in clinical trials.[11] Automated disease-specific image assessment tools would make objective measurements widely accessible to providers and patients for both research and, after sufficient validation, clinical decision-making. However, the capability to convert subjective qualitative expert judgment into quantitative values may prove
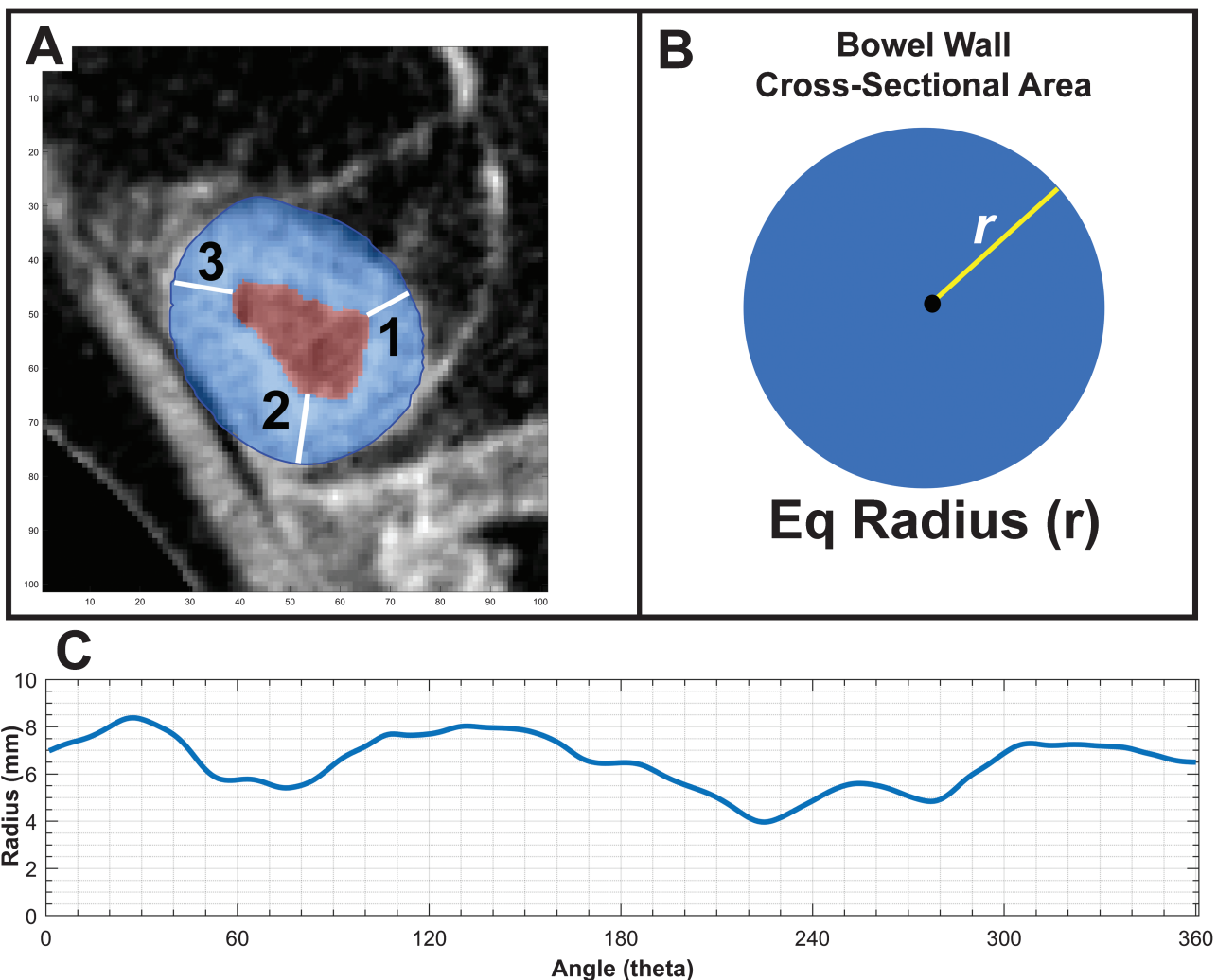

FIGURE 3. Comparing linear on-screen measurements and automated equivalent linear measures derived from cross-sectional areas. A, An example axial section of diseased small bowel with bowel wall masked in blue and lumen masked in red based on automated segmentation. B, The process of equivalent radius generation, where the total bowel wall area is reformed into a perfect circle and the (equivalent) radius is measured as the square root of Area/ π. In this case, the equivalent BWT radius is 7.3 mm compared with radiologist A measuring 8.1 mm and radiologist B measuring 7.1 mm. C, The variability of bowel wall thickness over 360 degrees of rotation about the estimated lumen center. Measures vary between a bowel wall thickness of 8.3 mm and 4.1 mm based on the location of measurement in the same cross-sectional segment.

to be an equally valuable offering of computational image analysis. Ulceration, signal intensity, and mural stratification are examples of qualitative features contained in some IBD image-based disease activity scoring systems.[19, 20] The quantification of qualitative bowel damage features, such as strictures, offers improved standardization of both definitions and severity grading of present subjectively classified findings.

Comparative studies examining small bowel segmentation in Crohn's disease are sparse, owing to the challenges presented by the irregular intestinal contours, lumen and bowel wall heterogeneity, and difference in image acquisition between scans. Using MRE, Naziroglu et al successfully demonstrated the potential for semi-automated bowel wall thickness measurements by way of centerline and active contour intestinal segmentation using an image processing tool set (3DNetSuite, Biotronics3D Inc., London, UK).[21] They reported BWT correlations between radiologists and automated measures in 27 patients with active small bowel CD that were similar to the agreement reported in our study, ranging from an ICC = 0.542 to an ICC = 0.737. Work from the same group used semi-automated bowel wall thickness and enhancement measurements in conjunction with subjective mural enhancement signal to produce a radiographic activity score good correlation and overall accuracy compared with the endoscopic scores.[22] Importantly, scores using semi-automated components had superior interobserver agreement (ICC = 0.81 vs ICC = 0.44–0.59) compared with other imaging activity scores, including the London and MaRIA scores. At the least, automated measures can immediately offer improved reproducibility with similar performance compared with existing scoring systems.

Though encouraging, the results should be considered in the context of several limitations. First, this study used a a retrospective design where imaging studies were deliberately selected to avoid confounders of poor image quality, severe penetrating features, and complex postsurgical anatomy. The results presented will require further validation before implemented into clinical practice. Additionally, although this work represents an incremental step toward fully automated measurements, human operators are still needed at present for placement of intestinal reference points and quality control. Measurement ground truth, or "gold-standards," are difficult to define. Radiologist A, B, or automated measures could each be closest to the actual intestinal distances and geometries. As referenced earlier, our paired radiologists' BWT correlation ($r = 0.724$) was similar to expert BWT correlation reported in work by others (ICC = 0.73 to 0.74). For this reason, we believe the expert-to-expert correlation is a reasonable benchmark for practically assessing the accuracy of semi-automated measurements pending direct comparisons to gross pathology.

Another important point of discussion is the method by which bowel measurements are collected. Manual assessments involve the radiologist measuring the minimum or maximum linear dimension typically using electronic rulers or calipers.

However, intestinal walls are often irregular, and linear measurements do not account for variation along the 360-degree rotational axis of the cylindrical shape of the intestine. These considerations contribute to skew or rotation of linear measurements—not to mention that smaller scale measurements can be challenging and magnify error between observers. The semi-automated measurements in this study assess incremental cross-sectional areas of bowel, allowing capture of all 360 degrees of intestinal features. Therefore, the derived equivalent radius or diameter generated by these automated techniques are fundamentally different and inherently more accurate than how reference radiologists measure bowel features (Fig. 3). Although existing measurements, evaluation tools, and scoring systems use many of the linear values studied in this work, new concepts in bowel measurement may prove to be additive or superior to human linear measures.

Finally, though promising, the presented semi-automated methods are far from a complete bowel damage assessment. Penetrating features were not included in this dataset, as existing segmentation technology is often challenged in identifying these complications. Additionally, extra-intestinal disease features, such as lymphadenopathy and local mesenteric hyperemia, are
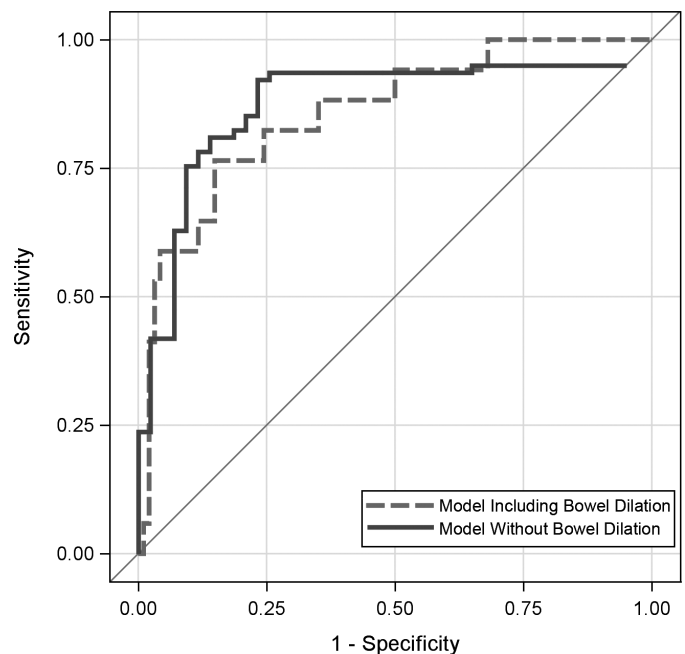

FIGURE 4. Performance of models predicting the presence of radiologist-defined intestinal strictures.Using automatically derived bowel measurements, models predicting radiologist-defined stricturing disease were generated. Using a stricture definition of 25% or more increase in bowel wall thickness and 50% or more reduction of lumen caliber relative to radiographically normal adjacent bowel and including criteria of 30 mm or more of upstream bowel dilation, automated models had an AuROC of 0.86 (dashed line). Alternatively, when removing the criteria for bowel dilation, model performance improved with an AuROC of 0.92 (solid line).

important but are unmeasured by the presented methods. This highlights the supplemental benefits of these automated methods to the radiologist's interpretation. Exposure to diagnostic radiation is a consideration, though in many countries CT is more readily available than MR-enterography, and increasingly low radiation protocols are substantially minimizing radiation-associated risks.[23] Having bowel damage assessment tools compatible with the technology available in most healthcare settings will help facilitate the personalization of care outside of high-volume academic centers.

CONCLUSION

In conclusion, semi-automated measurements of structural damage in Crohn's disease approach the correlation seen between paired experienced radiologists. Additionally, intestinal strictures, which are often qualitatively defined, can be quantitatively measured using image analysis tools. Like genomics, proteomics, and metabolomics, radiomics offers another layer for describing and understanding both phenotype and mechanisms of disease. A potential near future state is one where computational image analysis provides not only fully automated disease assessments but also additionally new insights, measurements, and predictions that are not possible using existing tools. Conceivable applications include therapeutic decision support tools, integration into telemedicine programs, and providing individual-level phenotyping for population-based studies. Ongoing work in IBD image analysis will include efforts to quantify disease activity features, link novel features to clinical outcomes, and—importantly—study the human-computer interaction between both providers and patients with emerging digital analytics technologies.

REFERENCES

1. Sturm A, Maaser C, Calabrese E, et al. ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 2: IBD scores and general principles and technical aspects. *J Crohns Colitis.* 2018;13:273–284.
2. Fiorino G, Morin M, Bonovas S, et al. Prevalence of bowel damage assessed by cross-sectional imaging in early Crohn's disease and its impact on disease outcome. *J Crohns Colitis.* 2017;11:274–280.
3. Rieder F, Fiocchi C, Rogler G. Mechanisms, management, and treatment of fibrosis in patients with inflammatory bowel diseases. *Gastroenterology.* 2017;152:340–350.e6.
4. Deepak P, Fletcher JG, Fidler JL, et al. Radiological response is associated with better long-term outcomes and is a potential treatment target in patients with small bowel Crohn's disease. *Am J Gastroenterol.* 2016;111:997–1006.
5. Deepak P, Fletcher JG, Fidler JL, et al. Predictors of durability of radiological response in patients with small bowel Crohn's disease. *Inflamm Bowel Dis.* 2018;24:1815–1825.
6. Bruining DH, Siddiki HA, Fletcher JG, et al. Benefit of computed tomography enterography in Crohn's disease: effects on patient management and physician level of confidence. *Inflamm Bowel Dis.* 2012;18:219–225.
7. Rimola J, Panés J, Ordás I. Magnetic resonance enterography in Crohn's disease: optimal use in clinical practice and clinical trials. *Scand J Gastroenterol.* 2015;50:66–73.
8. Danese S, Bonovas S, Lopez A, et al. Identification of endpoints for development of antifibrosis drugs for treatment of Crohn's disease. *Gastroenterology.* 2018;155:76–87.
9. Rieder F, Bettenworth D, Ma C, et al. An expert consensus to standardise definitions, diagnosis and treatment targets for anti-fibrotic stricture therapies in Crohn's disease. *Aliment Pharmacol Ther.* 2018;48:347–357.
10. Bruining DH, Zimmermann EM, Loftus EV Jr, et al.; Society of Abdominal Radiology Crohn's Disease-Focused Panel. Consensus recommendations for evaluation, interpretation, and utilization of computed tomography and magnetic resonance enterography in patients with small bowel Crohn's disease. *Gastroenterology.* 2018;154:1172–1194.
11. Jairath V, Ordas I, Zou G, et al. Reliability of measuring ileo-colonic disease activity in Crohn's disease by magnetic resonance enterography. *Inflamm Bowel Dis.* 2018;24:440–449.
12. Hou JK, Tan M, Stidham RW, et al. Accuracy of diagnostic codes for identifying patients with ulcerative colitis and Crohn's disease in the Veterans Affairs Health Care System. *Dig Dis Sci.* 2014;59:2406–2410.
13. Gryska EA, Schneiderman J, Heckemann RA. Automatic brain lesion segmentation on standard MRIs of the human head: a scoping review protocol. *BMJ Open.* 2019;9:e024824.
14. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542:115–118.
15. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al.; the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama.* 2017;318:2199–2210.
16. Pariente B, Mary JY, Danese S, et al. Development of the Lémann index to assess digestive tract damage in patients with Crohn's disease. *Gastroenterology.* 2015;148:52–63.e3.
17. Tielbeek JA, Makanyanga JC, Bipat S, et al. Grading Crohn disease activity with MRI: interobserver variability of MRI features, MRI scoring of severity, and correlation with Crohn disease endoscopic index of severity. *AJR Am J Roentgenol.* 2013;201:1220–1228.
18. Siddiki HA, Fidler JL, Fletcher JG, et al. Prospective comparison of state-of-the-art MR enterography and CT enterography in small-bowel Crohn's disease. *AJR Am J Roentgenol.* 2009;193:113–121.
19. Prezzi D, Bhatnagar G, Vega R, et al. Monitoring Crohn's disease during anti-TNF-α therapy: validation of the magnetic resonance enterography global score (MEGS) against a combined clinical reference standard. *Eur Radiol.* 2016;26:2107–2117.
20. Rimola J, Rodriguez S, García-Bosch O, et al. Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. *Gut.* 2009;58:1113–1120.
21. 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;90:20160654.
22. Puylaert CAJ, Schüffler PJ, Naziroglu RE, et al. Semiautomatic assessment of the terminal ileum and colon in patients with Crohn disease using MRI (the VIGOR++ Project). *Acad Radiol.* 2018;25:1038–1045.
23. Ippolito D, Lombardi S, Trattenero C, et al. CT enterography: diagnostic value of 4th generation iterative reconstruction algorithm in low dose studies in comparison with standard dose protocol for follow-up of patients with Crohn's disease. *Eur J Radiol.* 2016;85:268–273.