Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer

Abdominal Radiology (2020) 45:632–643
https://doi.org/10.1007/s00261-019-02321-8

PELVIS

Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer

Joost J. M. van Griethuysen[1,2] · Doenja M. J. Lambregts[1] · Stefano Trebeschi[1,2] · Max J. Lahaye[1] · Frans C. H. Bakers[3] · Roy F. A. Vliegen[4] · Geerard L. Beets[5,2] · Hugo J. W. L. Aerts[6,2] · Regina G. H. Beets-Tan[1,2]

Published online: 16 November 2019

Abstract

Purpose To compare the performance of advanced radiomics analysis to morphological assessment by expert radiologists to predict a good or complete response to chemoradiotherapy in rectal cancer using baseline staging MRI.

Materials and methods We retrospectively assessed the primary staging MRIs [prior to chemoradiotherapy (CRT)] of 133 rectal cancer patients from 2 centers. First, two expert radiologists subjectively estimated the likelihood of achieving a "complete response" (ypT0) and "good response" (TRG 1–2), using a 5-point score (based on TN-stage, MRF/EMVI-status, size/signal/shape). Next, tumor volumes were segmented on high *b* value DWI (semi-automated, corrected by 2 non-expert and 2-expert readers, resulting in 5 segmentations), copied to the remaining sequences after which a total of 2505 radiomic features were extracted from T2W, low and high *b* value DWI and ADC. Stability of features for noise due to inter-reader and inter-scanner and protocol variations was assessed using intraclass correlation (ICC) and the Kruskal–Wallis test. Using data from center 1 ($n=86$; training set), top 9 features were selected using minimum Redundancy Maximum Relevance and combined in a logistic regression model. Finally, diagnostic performance of the fitted models was assessed on data from center 2 ($n=47$; validation set) and compared to the performance of the radiologists.

Results The Radiomic models resulted in AUCs of 0.69–0.79 (with similar results for the segmentations performed by expert/non-expert readers) to predict response, results similar to the morphologic prediction by the expert radiologists (AUC 0.67–0.83). Radiomics using semi-automatically generated segmentations (without manual input) did not result in significant predictive performance.

Conclusions Radiomics could predict response to therapy with comparable diagnostic performance as expert radiologists, regardless of whether image segmentation was performed by non-expert or expert readers, indicating that expert input is not required in order for the radiomics workflow to produce significant predictive performance.

Keywords Rectal cancer · Magnetic resonance imaging · Response prediction · Radiomics · Texture analysis

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00261-019-02321-8) contains supplementary material, which is available to authorized users.

✉ Doenja M. J. Lambregts
d.lambregts@nki.nl

[1] Department of Radiology, The Netherlands Cancer Institute, PO Box 90203, 1006 BE Amsterdam, The Netherlands

[2] GROW School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands

[3] Department of Radiology & Nuclear Medicine, Maastricht University Medical Center, Maastricht, The Netherlands

[4] Department of Radiology, Zuyderland Medical Center, Heerlen, The Netherlands

[5] Department of Surgery, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[6] Department of Radiation Oncology and Radiology, Computational Imaging and Bioinformatics Laboratory, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA

Introduction

According to current standard of care, patients with very distal and/or locally advanced rectal tumors (≥ T3 and/or N+) typically receive neoadjuvant chemoradiotherapy (CRT) aiming to achieve downstaging and thereby increasing the chance of a complete surgical resection. As a result of CRT, approximately 15% of patients undergo a complete tumor response [1]. There is a current paradigm shift in treatment towards considering organ preservation ('watch-and-wait') for these very good responders [1–3]. In addition to assessing response after completion of CRT to select these patients, there is also an increased clinical interest in the prediction of treatment response before the start of CRT. In patients likely to respond well, neoadjuvant treatment may be intensified, for example with an additional radiotherapy boost, to increase the chance of organ preservation. Patients with smaller tumors have a higher response rate [4, 5], but according to current standards are typically treated with direct surgery without CRT. However, with a predicted high response rate chemoradiation might be offered to these small tumors as an alternative with the sole aim to achieve organ preservation, whereas patients with radioresistant tumors remain better off with surgery alone, which is the current standard treatment for these tumors. To date, such an approach is obviously still experimental and offered only in trial settings, for example within the STAR-TREC study, a collaborative phase II trial on CRT+organ preservation for early rectal cancer running in the UK, Denmark and the Netherlands (ClinicalTrials.gov NCT02945566) [6].

Several studies have shown that imaging may play a role in the pre-treatment prediction of response, with a particular focus on MRI being one of the main imaging modalities used to stage rectal cancer. "Semantic features" including the T-stage, N-stage, Circumferential Resection Margin (CRM), Extra-Mural Venous Invasion (EMVI) and baseline tumor volume have been shown to be associated with the chance of response to varying degrees [7–10]. Promising (though inconsistent) results have also been reported for the use of more novel functional MR imaging sequences such as diffusion-weighted imaging (DWI) and dynamic contrast enhanced (DCE) MRI, that can provide quantifiable information on biological tumor properties such as tumor cellularity and tumor perfusion [11–13].

Another highly interesting recent development is radiomics, a high-throughput post-processing technique capable of extracting large numbers of quantitative "features" from routinely acquired medical imaging [14]. These features can be used to generate a comprehensive radiologic phenotype and can potentially provide us with new insights into underlying biologic tumor characteristics [15–17]. In rectal cancer, a handful of studies investigating radiomics for response prediction have shown promising results [18–20], albeit mainly in relatively small single center cohorts. So far, no studies exist that have compared the use of radiomics to subjective estimation of the likelihood of response by radiologists based on an overall visual interpretation of the local tumor stage at baseline MRI. Such a comparison would be an interesting step to provide at least some preliminary perspective on the potential added benefit from radiomics in a clinical setting.

With this study we aim to add to previous research by investigating the potential of radiomics to predict treatment response in rectal cancer using the baseline staging MRI data from two institutions (to allow a test and validation dataset and to study effects of acquisition heterogeneity) and by comparing the performance of radiomics to morphological assessment of the images by expert radiologists to provide a first exploratory estimation of its potential clinical benefit.

Methods and materials

The study was approved by the local institutional review board (of both institutions). Due to the retrospective nature of the study, informed consent was waived.

Study population

We retrospectively identified 133 patients with rectal cancer who underwent long course chemoradiotherapy at one of two study centers (Maastricht University Medical Center and Zuyderland Medical Center Heerlen) between March 2007 and January 2013. Main inclusion criteria were (a) histologically proven primary non-mucinous type rectal adenocarcinoma, (b) locally advanced disease (≥ cT3 and/or N+ disease), (c) neoadjuvant treatment consisting of 28 fractions of 1.8 Gy radiotherapy with concurrent capecitabine 825 mg/m^2 chemotherapy, (c) availability of a multiparametric pre-treatment MR examination including a T2-weighted sequence, a diffusion-weighted sequence and corresponding quantitative 'apparent diffusion coefficient' (ADC) map, and (d) availability of either histology after surgery or long-term (> 2 years) follow-up in case of a wait-and-see program to establish the final treatment response.

Image acquisition

All patients received a primary staging MRI on a 1.5T MR system (Intera or Ingenia MR system; Philips Healthcare, Best, The Netherlands in center 1; Magnetom Avanto; Siemens in center 2).

The imaging protocol included a T2-weighted turbo spin echo sequence in sagittal, coronal and transverse plane and a transverse EPI-DWI sequence with 1000 or 1100 s/mm^2 as the highest *b* value. Detailed sequence parameters are provided in Table 1. ADC maps were calculated from the DWI sequences using a mono-exponential model including all available *b* values. Oblique transverse T2W and DWI sequences were acquired in identical planes perpendicular to the tumor axis as seen on the sagittal T2W scan. The transverse T2-weighted, low *b* value DWI (DWI_{b0}), high *b* value DWI ($DWI_{b1000/b1100}$) images and the ADC maps were used for radiomic feature extraction.

Standard of reference/clinical outcome

The main clinical study outcomes were:

(1) the prediction of a complete versus incomplete response after chemoradiotherapy.
(2) the prediction of a good versus poor response after chemoradiotherapy.

The final histopathologic tumor stage after surgery including the tumor regression grade (TRG) according to Mandard [21] served as the main standard of reference. For the first study outcome, patients with a ypT0/TRG1 were classified 'complete responders,' while patients with residual tumor (ypT1-4, TRG 2–4) were classified as 'incomplete responders.' For the second outcome, patients with a TRG1-2 (indicating predominant fibrosis) were classified as 'good responders' and patients with TRG3-5 as 'poor responders'. For $N = 13$ patients who underwent wait-and-see without surgery, a sustained clinical complete response for > 2 years follow-up (i.e., no signs of recurrence on follow-up MRI and endoscopy performed 3 monthly in the first year and 6-monthly in the following years) was used as a surrogate endpoint for a complete response. These patients were included in the 'complete response' and 'good response' groups for the two respective outcomes.

Visual morphologic assessment by expert radiologists

Two independent board-certified abdominal radiologists (DMJL and MJL), with each > 10 years' specific experience in reading rectal MRI, estimated the likelihood of whether a patient would achieve a complete response (outcome 1) or good response (outcome 2), respectively, using a 5-point subjective confidence score (1 = chance to achieve a complete/good response highly unlikely, 2 = good/complete response unlikely, 3 = equivocal, 4 = complete/good response likely, 5 = complete/good response highly likely). The readers based their score on their overall visual morphologic assessment of the size, signal and shape of the tumor, T- and N-stage, circumferential resection margin (CRM) and EMVI, according to the criteria described in Table 2. Tumors with more unfavorable characteristics (e.g., larger size, higher T-stage, positive N-stage, CRM +, EMVI +) were assigned lower scores. Readers were blinded for the patient's outcome and each other's results.

Table 1 MR acquisition protocols

	Diffusion-weighted MRI					T2 weighted MRI	
	Center 1			Center 2		Center 1	Center 2
	DWI 1	DWI 2	DWI 3	DWI 1	DWI 2	T2W	T2W
N slices	50	24	20–24	34	34	22–34	48–60
Repetition time	3969–5503	3731–5545	4141–5240	5100	4300–5314	3378–9557	3400–4670
Echo time	70	70–73	65–70	88	79	130–150	118–122
Flip angle	90	70	70	90	90	90	150
Phase encoding direction	AP	AP	AP	AP	AP	LR	LR
In-plane spacing (mm × mm)	1.7 × 1.7	1.25 × 1.25	1.25 × 1.25	1.25 × 1.25	2.0 × 2.0	0.8 × 0.8–0.4 × 0.4	0.8 × 0.8
Slice thickness	5	5	5	5	6	3–5	3.5
Echo train length	1	1	1	1	1	25–26	23
NSA	4–10	3–5	5	6	6	2–6	2
Fat saturation	STIR	SPIR	SPAIR	SPIR	SPIR	N/A	N/A
EPI factor	47–55	55–77	61–83	148	150	N/A	N/A
b values	0, 500, 1000	0, 500, 1000	0, (25, 50, 100), 500, 1000	0,500,1000	0,300,1100	N/A	N/A

AP anterior–posterior, *LR* left–right, *NSA* number of signals averaged, *EPI* echo planar imaging, *STIR* short TI inversion recovery, *SPIR* spectral presaturation inversion recovery, *SPAIR* spectral attenuated inversion recovery

Table 2 Likelihood score used by the two radiologists to predict the chance of achieving a good or complete response, respectively, based on visual evaluation of the baseline MRI

Likelihood of achieving outcome	1: Highly unlikely	2: Unlikely	3: Equivocal	4: Likely	5: Highly likely
Criteria					
Size	Large (> 5 cm)	Large (> 5 cm)	–	Small (< 3 cm)	Small (< 3 cm)
Signal	Heterogeneous	Heterogeneous	–	Homogenous	Homogenous
T-stage	≥ T3 cd	≥ T3 cd	–	≤ T3ab	≤ T3ab
Shape	Irregular	Irregular	–	Regular	Regular
N-stage	N+	N+	–	N0	N0
EMVI	EMVI+	EMVI+	–	EMVI-	EMVI-
CRM	CRM+	CRM+	–	CRM-	CRM-
Outcome					
Good response	All 7 criteria	≥ 5 criteria	Not meeting the criteria for scores 1–2 or 4–5	≥ 3 criteria	≥ 5 criteria
Complete response	≥ 5 criteria	≥ 3 criteria		≥ 5 criteria	all 7 criteria

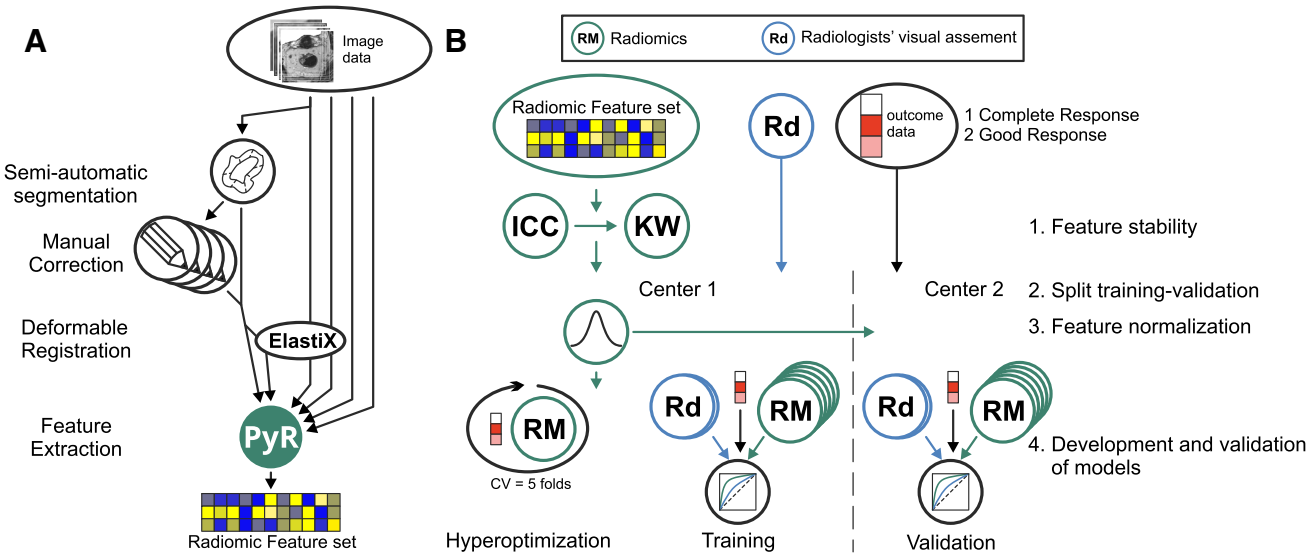

Fig. 1 Study workflow describing the image segmentation, registration and radiomic feature extraction steps (**a**) and data analysis steps (**b**). **a** Tumors were segmented on DWI_{b1000}. After co-registration of T2W and DWI_{b0} images, segmentations were transformed for extraction from T2W images using the transformation map from the registration. Images and segmentation maps were then fed into the PyRadiomics pipeline (PyR) for feature extraction. **b** After exclusion of unstable features (1), data were divided into training and validation sets by center, with center 1 used for training and 2 for validation (2). Feature values were normalized using mean and standard deviation of features in center 1 (3). Using the training set and 5-fold stratified cross-validation, optimal hyperparameters were determined for the radiomics model. The optimized model was then trained on the full training set for each reader and each outcome separately. Finally, performance to predict response was assessed in the training and validation sets (4). *ICC* intraclass correlation coefficient, *KW* Kruskal–Wallis, *CV* cross-validation

Radiomics workflow

The radiomics feature extraction workflow, including image segmentation and radiomic feature extraction as the two main steps, is schematically illustrated in Fig. 1a.

I—Image segmentation

The image segmentation comprises the first 3 steps of the workflow:

Step 1: Semi-automatic segmentation

Tumor volumes were semi-automatically segmented on the high *b* value diffusion images using a region-growing algorithm implemented in MANGO (Multi-image Analysis GUI, version 3.8, Research Imaging Institute, University of Texas Health Science Center, San Antonio, TX), according to methods previously reported [22]. The high *b* value images were chosen as they provide a good tumor-to-background signal ratio.

Step 2: Manual adjustment

The tumor segmentations derived in step 1 were then checked and manually adjusted where deemed necessary by four independent readers (two resident level non-expert readers (JJMVG and ST) and two expert radiologists (DMJL and MJL)) to allow assessment of effects of interobserver variations and reader experience level (see Fig. 2). This resulted in a total of 5 segmentations (1 semi-automated, 2 non-expert, 2 expert) used for radiomic feature extraction. Overlap between segmentations was assessed using the dice similarity coefficient.

Step 3: Registration of different imaging sequences

To correct for organ displacements and deformations, T2W and DWI images were co-registered using deformable B-spline registration implemented in Elastix [23, 24]. The resulting deformation maps were then used to adapt the DWI-based segmentations to the T2W images.

II—Radiomic feature extraction

Feature extraction was performed using the PyRadiomics toolbox (version 2.1.2) [25]. Prior to feature extraction, images were normalized to 0 mean and 100 standard deviation to reduce influence of differences in MR system vendor and acquisition protocol between the two centers [26] and subsequently interpolated to isotropic voxels with 2 mm sides using a B-Spline interpolator. To remove outlier intensity values, the five segmentations were resegmented by excluding voxels which differed $> 3\sigma$ from the mean. Prior to extraction of texture features and first order Uniformity and Entropy, gray values were discretized using a fixed bin width of 5. For each sequence (T2W, DWI_{b0}, $DWI_{b1000/11000}$, ADC), 623 intensity and texture features were extracted from non-derived, gradient, exponent, logarithm and Laplacian of Gaussian ($\sigma \in \{1\text{ mm}, 3\text{ mm}, 5\text{ mm}\}$) filtered images (yielding $4 \times 623 = 2492$ features). In addition, 13 shape descriptors were extracted from non-resegmented DWI-based segmentations, resulting in a grand total of 2505 ($2492 + 13$) features for each of the five respective segmentations (1 semi-automated, 2 non-expert readers and 2 expert readers). The PyRadiomics configuration file used is provided in the supplementary materials.

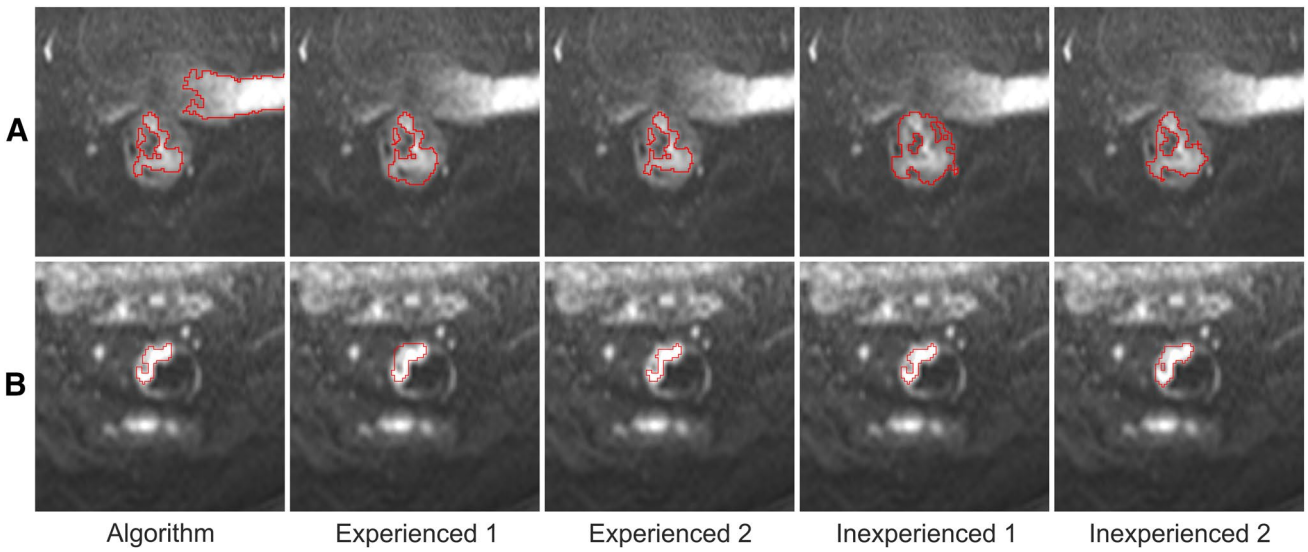

Fig. 2 Example of results of semi-automatic (algorithm) tumor segmentation performed on high *b* value diffusion-weighted images and results after manual correction by two expert and two non-expert readers. **a** Example of a case where the semi-automatic algorithm erroneously included a high signal band-shaped artifact in the tumor segmentation, which was discarded after manual adjustment by both non-expert and expert readers. **b** Example of a case where the semi-automated segmentation by the algorithm performed well and did not require significant changes by either non-expert or expert readers

Statistical analysis

Statistical analysis was performed using the Python (v3.5.3) package Scikit-learn (v0.20) [27] and is schematically illustrated in Fig. 1b. Data from center 1 ($n = 86$) were used for training, data from center 2 ($n = 47$) were used for validation. Baseline characteristics were analyzed using χ^2-test for categorical variables and independent samples t test for continuous variables. Stability of radiomic features for inter-reader variation was assessed using intraclass correlation coefficient (ICC) and stability for differences in MR system vendor and acquisition protocol between the 2 centers was assessed using the Kruskal–Wallis (KW) test. Only features exhibiting sufficient stability (ICC ≥ 0.75 and KW p value ≥ 0.05) were eligible for selection in the radiomics prediction model. Stable features were normalized by subtracting the mean and dividing by the standard deviation on a per-reader basis. Mean and standard deviation for each feature is determined using only data from center 1 (training set).

Using the training set, the radiomics model was trained separately for each of the 5 segmentations in 2 steps: (1) Using minimum Redundancy Maximum Relevance (mRMR), implemented in Python package 'mifs' [28], a set of candidate features was selected from the training set, which (2) were fitted into a logistic regression model with l2 regularization and balanced class weights. To approximate the mutual information between the outcome and continuous features during mRMR selection, we employed the nearest neighbor method as described by Ross et al. [29] Optimum number of features to select [5–10], as well as the k neighbors parameter [5–8] in mRMR and the C regularization parameter [10^{-7}–10^{2}] in the logistic regression model were determined by 5-fold stratified cross-validation on the training set. Finally, the performance of the radiomics model to predict a 'complete' and 'good' response, respectively, was assessed using the Wilcoxon rank-sum test and by calculating the area under the ROC curve (AUC). Using the DeLong method [30], AUC for radiomics was then compared to the AUC calculated for the morphologic prediction of response by the two expert radiologists based on their subjective confidence scores. p values < 0.05 were considered significant. Interobserver agreement for the subjective scoring by the two radiologists was assessed using quadratic Cohen's kappa.

Results

Baseline characteristics

The baseline characteristics of the patients are shown in Table 3. No significant differences were seen between the two centers. In total 28 patients were complete responders (15 after surgery and 13 sustained clinical complete responders undergoing W&S) and 105 patients had residual tumor. For the second clinical outcome, good versus poor response, 62 patients were considered good responders (28 TRG1, 34 TRG2) and 67 poor responders (38 TRG3, 25 TRG4, 4 TRG5). In 4 patients (3 from center 1, 1 from center 2) no TRG stage was available, these patients were therefore excluded from the latter analysis.

Performance of radiologists' visual morphologic assessment to predict response

Results for the prediction of response by the two radiologists (compared to the performance of the radiomics models) are provided in Table 4 and illustrated in Fig. 3. Overall, AUC to predict a complete response in the validation cohort was 0.83 for the first reader and 0.74 for the second reader. For the prediction of a good response, AUC was 0.68 (reader 1) and 0.67 (reader 2) in the validation cohort. Agreement between the two readers was good with $\kappa = 0.64$ and $\kappa = 0.61$ for the prediction of complete and good response, respectively.

Building the radiomics models

1692 out of the in total 2505 radiomic features (68%) showed an ICC ≥ 0.75 (indicating sufficient inter-reader stability), of which only 415 (25%) showed no confounding related to the MR system and acquisition protocol used (i.e., MRI performed in center 1 or center 2). These 415 features were considered stable and available for selection by the radiomics model. Optimum settings for the model, as determined by the hyper-optimization in the training set, turned out to be 9 features, $k = 8$ neighbors and $C = 10^{-5}$. Most emphasis was placed on DWI and ADC sequences, with only few features selected from T2W sequences in 8/10 developed models. Further details regarding the selected features per model are provided in Supplementary materials 2.

Table 3 Baseline characteristics

	Total ($N=133$)	Center 1 ($N=86$)	Center 2 ($N=47$)	P value
Age	68 [45–87]	69 [48–87]	67 [45–85]	0.1583[a]
Gender				0.691[b]
M	92 (69.2%)	61 (70.9%)	31 (66.0%)	
F	41 (30.8%)	25 (29.1%)	16 (34.0%)	
Initial cT stage before treatment				0.931[b]
1–2	16 (12.0%)	11 (12.8%)	5 (10.6%)	
3	109 (82.0%)	70 (81.4%)	39 (83.0%)	
4	8 (6.0%)	5 (5.8%)	3 (6.4%)	
Initial cN stage before treatment				0.316[b]
0	11 (8.3%)	8 (9.3%)	3 (6.4%)	
1	40 (30.1%)	22 (25.6%)	18 (38.3%)	
2	81 (60.7%)	56 (65.1%)	26 (55.3%)	
Final treatment after CRT				0.065[b]
TME	119 (89.5%)	73 (84.9%)	46 (97.9%)	
W&W	13 (9.8%)	12 (14.0%)	1 (2.1%)	
TEM	1 (0.7%)	1 (1.1%)	0 (0%)	
Final yT stage				0.917[b]
0	28 (21.0%)	18 (20.9%)	10 (21.3%)	
1	11 (8.3%)	8 (9.3%)	3 (6.4%)	
2	30 (22.6%)	19 (22.1%)	11 (23.4%)	
3	59 (44.4%)	37 (43.0%)	22 (46.8%)	
4	5 (3.8%)	4 (4.7%)	1 (2.1%)	
Final yN stage				0.233[b]
0	94 (70.7%)	65 (75.6%)	29 (61.7%)	
1	29 (21.8%)	16 (18.6%)	13 (27.7%)	
2	10 (7.5%)	5 (5.8%)	5 (10.6%)	
Tumor regression grade (TRG)				0.108[b]
1	28 (21.0%)	18 (20.9%)	10 (21.3%)	
2	34 (25.6%)	24 (27.9%)	10 (21.3%)	
3	38 (28.6%)	26 (30.2%)	12 (25.5%)	
4	25 (18.8%)	11 (12.8%)	14 (29.8%)	
5	4 (3.0%)	4 (4.7%)	0 (0%)	
Missing	4 (3.0%)	3 (3.5%)	1 (2.1%)	

TME total mesorectal excision, *W&W* wait & wait (organ saving treatment), *TEM* transanal endoscopic microsurgery, *cT, cN*: clinical T- and N-stage as assessed on primary MRI, *yT, yN* final T- and N-stage after nCRT as assessed at histopathology after surgery ($n=120$) or by long-term follow-up in case of wait-and-see treatment ($n=13$), *TRG* tumor regression grade

[a] t test

[b] χ^2 test

Performance of the radiomics models to predict response

In the training set, radiomics models based upon manually corrected segmentations showed significant performance to predict both 'complete response' (AUC 0.71 to 0.74) and 'good response' (AUC 0.69 to 0.77). In the validation dataset, AUCS ranged between 0.69 and 0.79 ($p=0.001$–0.028) to predict a good response, with comparable performance for the segmentations performed by the two non-expert and expert readers. For the prediction of complete response, only the radiomics model using the segmentations from 1 expert and 1 non-expert reader retained significant performance, with respective AUCs of 0.77 (p value 0.010) and 0.73 (p value 0.029). Performance of the radiomics model using the semi-automated segmentations (without manual reader input) was non-significant for both study outcomes. Average dice coefficients between the semi-automated and different non-expert and expert manual-input segmentations are shown in Table 5.

Table 4 Performance to predict response

Reader	Center 1 (training, $n=86$)			Center 2 (validation, $n=47$)		
	AUC (95%CI)	Statistic	p value	AUC (95% CI)	Statistic	p value
Complete response						
Morphologic assessment (R1)	0.77 [0.62–0.91]	3.503	< 0.001	0.83 [0.69–0.98]	3.197	0.001
Morphologic assessment (R2)	0.67 [0.51–0.84]	2.272	0.023	0.74 [0.58–0.90]	2.326	0.020
Radiomics (exp_segm1)	0.71 [0.57–0.86]	2.771	0.006	0.77 [0.58–0.96]	2.573	0.010
Radiomics (exp_segm2)	0.74 [0.60–0.88]	3.089	0.002	0.69 [0.47–0.91]	1.820	0.069
Radiomics (non-exp_segm1)	0.71 [0.55–0.86]	2.707	0.007	0.73 [0.51–0.94]	2.183	0.029
Radiomics (non-exp_segm2)	0.74 [0.59–0.88]	3.100	0.002	0.66 [0.42–0.89]	1.508	0.132
Radiomics (semi-aut_segm)	0.73 [0.60–0.86]	3.025	0.002	0.63 [0.42–0.84]	1.248	0.212
Good response						
Morphologic assessment (R1)	0.60 [0.49–0.72]	1.608	0.108	0.68 [0.53–0.83]	2.072	0.038
Morphologic assessment (R2)	0.68 [0.56–0.79]	2.805	0.005	0.67 [0.52–0.83]	2.005	0.045
Radiomics (exp_segm1)	0.77 [0.67–0.87]	4.235	< 0.001	0.79 [0.66–0.93]	3.368	0.001
Radiomics (exp_segm2)	0.69 [0.57–0.80]	2.933	0.003	0.69 [0.52–0.86]	2.194	0.028
Radiomics (non-exp_segm1)	0.72 [0.61–0.83]	3.488	< 0.001	0.78 [0.64–0.92]	3.235	0.001
Radiomics (non-exp_segm2)	0.71 [0.60–0.82]	3.361	0.001	0.70 [0.53–0.86]	2.260	0.024
Radiomics (semi-aut_segm)	0.65 [0.53–0.77]	2.377	0.017	0.60 [0.43–0.78]	1.197	0.231

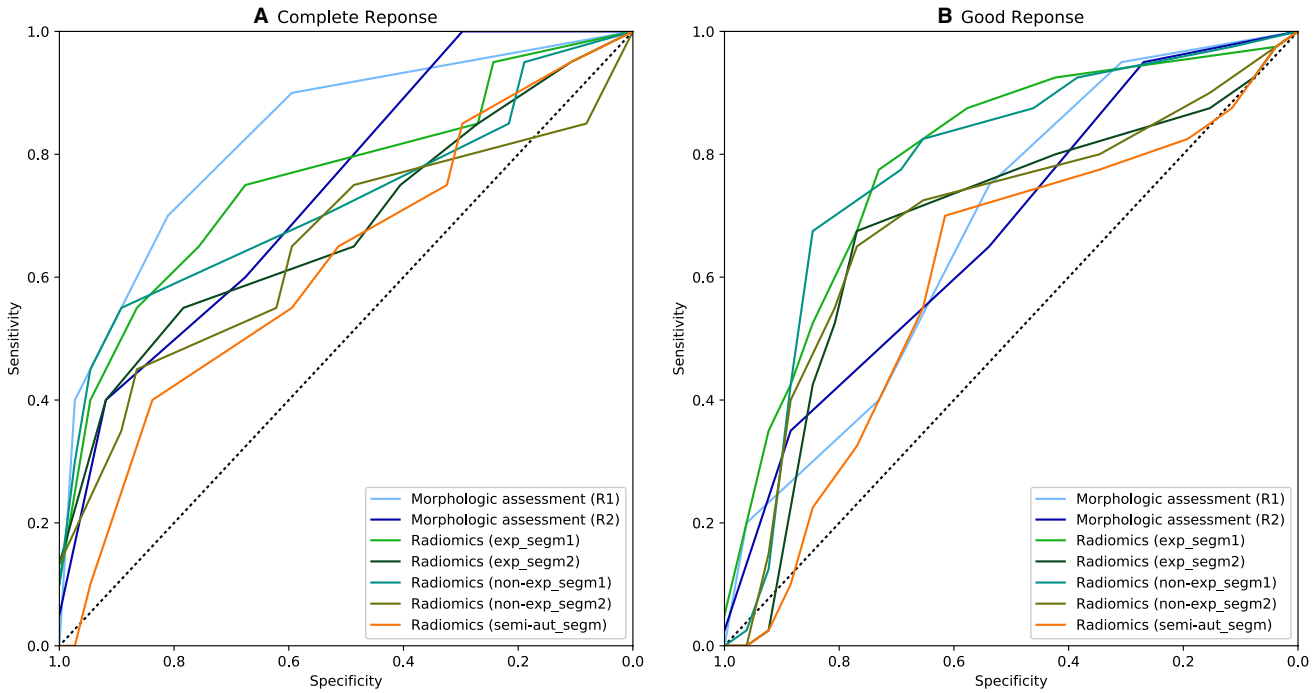

Fig. 3 ROC curves of morphologic assessment by radiologists and radiomics models to predict the outcome 'complete' (**a**) and 'good' (**b**) response. There were no statistically significant differences in diagnostic performance between the 2 radiologist readers and the various radiomics models

Comparison between radiomics model and morphologic assessment by radiologists

AUCs for the radiomics models (using segmentations with manual input) were comparable to the AUCs for the visual morphologic assessment by the expert radiologists, both for the prediction of a complete response (0.73–0.77 vs. AUC 0.74–0.83, $P=0.25$–0.88), as well as for the prediction of a good response (AUC 0.69–0.79 vs. AUC 0.67–0.68, $P=0.18$-0.93).

Table 5 Average (±SD) dice similarity between reader's segmentations

	exp_segm1	exp_segm2	non-exp_segm1	non-exp_segm2	semi-aut_segm
exp_segm1					
exp_segm2	0.84 (0.15)				
non-exp_segm1	0.73 (0.17)	0.71 (0.19)			
non-exp_segm2	0.77 (0.16)	0.84 (0.16)	0.69 (0.21)		
semi-aut_segm	0.78 (0.22)	0.76 (0.21)	0.64 (0.24)	0.78 (0.21)	

Discussion

In the present study, we compared the performance of advanced radiomics analysis to visual morphologic assessment by experienced radiologists to predict response to neoadjuvant chemoradiotherapy on primary staging MRI. Results show that the radiomics model could predict a good response to therapy upfront with similar diagnostic performance (AUC 0.69–0.79) as highly expert radiologists (AUC 0.67–0.68). Interestingly, the radiomics models were mainly based on features derived from DWI and ADC, with only few features selected from T2W imaging. This would suggest that DWI plays an important role when building response prediction models based on radiomics. Moreover, results of the radiomics model were comparable regardless of whether image segmentation was manually adapted by non-expert (young resident level) readers or by experienced radiologists, indicating that expert input is not required in order for the radiomics workflow to produce significant predictive performance. Radiomics models without manual input (using only semi-automated tumor segmentations) did not result in significant predictive performance, despite the fact that the spatial overlap between the semi-automated and manual-input segmentations was quite substantial (Dice 0.64–0.78).

Although in recent years several groups have investigated the potential of radiomics for rectal tumor response assessment, our current report is one of few to compare radiomics results to visual radiological assessment in order to put things into a more clinical perspective. To the best of our knowledge, only one previous report by Horvat et al. [31] compared performance of radiomics to expert reader assessment, though this study focused on response assessment after completion of therapy, rather than prediction upfront. In this study 34 features were extracted from 114 patients, and combined using a random forest classifier. This model showed excellent performance (AUC 0.93) in repeated cross-validation, which was significantly better than consensus scoring by 2 radiologists. A handful of previous studies specifically focused on MR-based radiomics to predict rectal tumor response prior to the start of treatment using baseline imaging data. Nie et al. [32] extracted 103 features from primary T1/T2, DWI and dynamic contrast enhanced MRI in 48 patients. Here, an artificial neural network was trained using 4-fold cross-validation to address overfitting, with resulting AUCs of 0.84 and 0.89 to predict a complete and good response, respectively. Cusumano et al. [20] performed a similar study but included an independent validation data cohort from another center. Here, a combination of shape, fractal and LoG-based features were assessed in a cohort of 198 patients, resulting in AUC 0.77 and 0.79 in the training and validation dataset, respectively. Finally, Cui et al. [19] assessed performance of radiomic features in 186 patients, achieving a very high AUC of 0.98 in the validation set. However, feature stability for image acquisition variation was not assessed, which may limit clinical applicability. In our current bi-institutional study we used a test and validation dataset from two independent centers in order to investigate potential confounding effects of variations in MR system vendor and acquisition protocols. We found that a large portion (75%) of features were classified as unstable to variations in vendor and image acquisition protocols. This highlights the need for standardized protocols and the importance of assessing feature stability when developing

radiomics models. On the other hand, despite these vendor and protocol variations, the radiomics models still achieved performance comparable to expert reader assessment.

Although, the AUCs of 0.66–0.79 achieved in our current study to predict response are encouraging, they will probably not yet be considered good enough for clinical decision making. As discussed above there is however still room for improvement. Further research should focus on standardization, but also on combining radiomic features with for example other clinical, histopathological, immunohistochemical or genetic biomarkers, which is likely to increase the predictive power, as has also been suggested by previous research [33, 34]. An accurate prediction of treatment response upfront, using biomarkers that can already be derived at baseline could impact clinical management in rectal cancer in the future. After completion of CRT, complete responders may already be accurately detected using a combination of simple visual DWI-MRI analysis and endoscopy, which limits the need for advanced imaging analysis tools such as radiomics in this setting [35]. Tools to predict treatment effects upfront are however not yet available in clinical practice. In locally advanced tumors, where downsizing is desired, but standard CRT is predicted to have little or no effect, one could consider a more intensified regimen or one that relies more on systemic therapy. If lateral resection margins are wide on MRI, one can even consider omitting neoadjuvant therapy altogether, avoiding unnecessary toxicity. In smaller rectal tumors, which are traditionally treated with TME surgery without neoadjuvant therapy, but which also have a higher chance to respond well to radiotherapy, a predictive model can guide treatment decisions towards (chemo)radiotherapy for the predicted responders with the goal to achieve organ preservation.

Our study design contained some limitations. The 95% confidence intervals for the performance of the radiological assessments as well as the radiomics models were large, most likely due to the relatively small size of the dataset, especially the validation set. Moreover, this small size of the training set can make the radiomics models prone to overfitting, as reflected by the fact that the optimum hyperparameters, with a low value for the C parameter and high value for the k parameter, favor high regularization. Our result will therefore need to be further validated in larger and preferably multicenter cohorts to obtain more stable results. Additionally, the estimated likelihood of achieving a good/complete response by the radiologists remains relatively subjective (despite the criteria provided in Table 2) and is dependent on the experience level of the two readers. We chose this approach to provide some preliminary perspective on how advanced model-based prediction methods would compare to what can potentially be achieved by mere "human" interpretation. We, however, acknowledge that an alternative approach including separate assessment of individual semantic features may allow for better reproducibility. This is a strategy we aim to further explore in future research. Finally, the analyses were all performed in patients with locally advanced rectal tumors. Our results will also need to be tested and validated in smaller tumors before radiomics models can be applied in these cases.

In conclusion, we were able to train radiomics models to reach comparable performance to predict response to chemoradiotherapy on baseline MRI as visual morphologic assessment and staging by highly expert radiologists, even when using tumor segmentations without any expert radiologist input. Furthermore, these results were obtained despite training on a very heterogeneous dataset, where the majority of features had to be excluded due to susceptibility for variations in image acquisition. Although validation in a large multicenter cohort is obviously needed, these results indicate that radiomics has strong potential to identify meaningful imaging biomarkers that can be included in clinically usable prediction models with the ultimate aim to further optimize and personalize treatment in rectal cancer.

Funding This work was supported by the Dutch Cancer Society (KWF) (Grant No. 2016-2/10611).

Compliance with ethical standards

Conflict of interest Dr. Aerts declares stock options in Sphera and Genospace, all other authors declare no conflict of interest.

References

1. Maas M, Beets-Tan RG, Lambregts DM, Lammering G, Nelemans PJ, Engelen SM, et al. (2011) Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. J Clin Oncol 29:4633–40. https://doi.org/10.1200/jco.2011.37.7176.
2. Martens MH, Maas M, Heijnen LA, Lambregts DMJ, Leijtens JWA, Stassen LPS, et al. (2016) Long-term Outcome of an Organ Preservation Program After Neoadjuvant Treatment for Rectal Cancer. J Natl Cancer Inst 108:djw171. https://doi.org/10.1093/jnci/djw171.
3. van der Valk MJM, Hilling DE, Bastiaannet E, Meershoek-Klein Kranenbarg E, Beets GL, Figueiredo NL, et al. (2018) Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the International Watch & Wait Database (IWWD): an international multicentre registry study. Lancet 391:2537–45. https://doi.org/10.1016/s0140-6736(18)31078-x.
4. Verseveld M, De Graaf EJR, Verhoef C, van Meerten E, Punt CJA, de Hingh IHJT, et al. (2015) Chemoradiation therapy for rectal cancer in the distal rectum followed by organ-sparing transanal endoscopic microsurgery (CARTS study). Br J Surg 102:853–60. https://doi.org/10.1002/bjs.9809.
5. Bujko K, Richter P, Smith FM, Polkowski W, Szczepkowski M, Rutkowski A, et al. (2013) Preoperative radiotherapy and local excision of rectal cancer with immediate radical re-operation for poor responders: a prospective multicentre study. Radiother Oncol 106:198–205. https://doi.org/10.1016/j.radonc.2012.12.005.

6. Rombouts AJM, Al-Najami I, Abbott NL, Appelt A, Baatrup G, Bach S, et al. (2017) Can we Save the rectum by watchful waiting or T rans A nal microsurgery following (chemo) R adiotherapy versus T otal mesorectal excision for early RE ctal C ancer (STAR-TREC study)?: protocol for a multicentre, randomised feasibility study. BMJ Open 7:e019474. https://doi.org/10.1136/bmjopen-2017-019474.
7. Maas M, Nelemans PJ, Valentini V, Das P, Rodel C, Kuo LJ, et al. (2010) Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. Lancet Oncol 11:835–44.
8. Curvo-Semedo L, Lambregts DMJ, Maas M, Thywissen T, Mehsen RT, Lammering G, et al. (2011) Rectal Cancer: Assessment of Complete Response to Preoperative Combined Radiation Therapy with Chemotherapy—Conventional MR Volumetry versus Diffusion-weighted MR Imaging. Radiology 260:734–43. https://doi.org/10.1148/radiol.11102467.
9. Lambregts DM, Rao SX, Sassen S, Martens MH, Heijnen LA, Buijsen J, et al. (2015) MRI and Diffusion-weighted MRI Volumetry for Identification of Complete Tumor Responders After Preoperative Chemoradiotherapy in Patients With Rectal Cancer: A Bi-institutional Validation Study. Ann Surg 262:1034–9. https://doi.org/10.1097/sla.0000000000000909.
10. Mahadevan LS, Zhong J, Venkatesulu B, Kaur H, Bhide S, Minsky B, et al. (2018) Imaging predictors of treatment outcomes in rectal cancer: An overview. Crit Rev Oncol Hematol 129:153–62. https://doi.org/10.1016/j.critrevonc.2018.06.009.
11. Hötker AM, Tarlinton L, Mazaheri Y, Woo KM, Gönen M, Saltz LB, et al. (2016) Multiparametric MRI in the assessment of response of rectal cancer to neoadjuvant chemoradiotherapy: A comparison of morphological, volumetric and functional MRI parameters. Eur Radiol:1–10. https://doi.org/10.1007/s00330-016-4283-9.
12. Martens MH, Subhani S, Heijnen LA, Lambregts DM, Buijsen J, Maas M, et al. (2015) Can perfusion MRI predict response to preoperative treatment in rectal cancer? Radiother Oncol 114:218–23. https://doi.org/10.1016/j.radonc.2014.11.044.
13. Chen Y-G, Chen M-Q, Guo Y-Y, Li S-C, Wu J-X, Xu B-H. (2016) Apparent Diffusion Coefficient Predicts Pathology Complete Response of Rectal Cancer Treated with Neoadjuvant Chemoradiotherapy. PLoS One 11:e0153944. https://doi.org/10.1371/journal.pone.0153944.
14. Hosny A, Parmar C, Quackenbush J, Schwartz LH, W L Aerts HJ. (2018). Artificial intelligence in radiology. https://doi.org/10.1038/s41568-018-0016-5.
15. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. (2012) Radiomics: the process and the challenges. Magn Reson Imaging 30:1234–48. https://doi.org/10.1016/j.mri.2012.06.010.
16. Gillies RJ, Kinahan PE, Hricak H. (2015) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:151169. https://doi.org/10.1148/radiol.2015151169.
17. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006. https://doi.org/10.1038/ncomms5006.
18. Liu Z, Zhang X-Y, Shi Y-J, Wang L, Zhu H-T, Tang Z-C, et al. (2017) Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. Clin Cancer Res:clincanres.1038.2017. https://doi.org/10.1158/1078-0432.ccr-17-1038.
19. Cui Y, Yang X, Shi Z, Yang Z, Du X, Zhao Z, et al. (2019) Radiomics analysis of multiparametric MRI for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Eur Radiol 29:1211–20. https://doi.org/10.1007/s00330-018-5683-9.
20. Cusumano D, Dinapoli N, Luca Boldrini ·, Chiloiro G, Gatta · Roberto, Masciocchi C, et al. (2018) Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. Radiol Med 123:286–95. https://doi.org/10.1007/s11547-017-0838-3.
21. Mandard AM, Dalibard F, Mandard JC, Marnay J, Henry-Amar M, Petiot JF, et al. (1994) Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. Cancer 73:2680–6.
22. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, Oei S, Rao S-X, de Graaff CAM, et al. (2016) Automated and Semi-automated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? Int J Radiat Oncol 94:824–31. https://doi.org/10.1016/j.ijrobp.2015.12.017.
23. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. (2010) Elastix: A toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 29:196–205. https://doi.org/10.1109/tmi.2009.2035616.
24. Shamonin DP, Bron EE, Lelieveldt BPF, Smits M, Klein S, Staring M, et al. (2014) Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease for the Alzheimer's Disease Neuroimaging Initiative. https://doi.org/10.3389/fninf.2013.00050.
25. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. (2017) Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77:e104–7. https://doi.org/10.1158/0008-5472.can-17-0339.
26. Collewet G, Strzelecki M, Mariette F. (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 22:81–91. https://doi.org/10.1016/j.mri.2003.09.001.
27. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIER-GRISEL O, Blondel M, Prettenhofer P, Weiss R, et al. (2011). Scikit-learn: Machine Learning in Python. vol. 12.
28. Hanchuan Peng, Fuhui Long, Ding C. (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–38. https://doi.org/10.1109/tpami.2005.159.
29. Ross BC. (2014) Mutual Information between Discrete and Continuous Data Sets. PLoS One 9:e87357. https://doi.org/10.1371/journal.pone.0087357.
30. DeLong ER, DeLong DM, Clarke-Pearson DL. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–45.
31. Horvat N, Veeraraghavan H, Khan M, Blazic I, Zheng J, Capanu M, et al. (2018) MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. Radiology 287:172300. https://doi.org/10.1148/radiol.2018172300.
32. Nie K, Shi L, Chen Q, Hu X, Jabbour S, Yue N, et al. (2016) Rectal Cancer: Assessment of Neoadjuvant Chemo-Radiation Outcome Based on Radiomics of Multi-Parametric MRI. Clin Cancer Res. https://doi.org/10.1158/1078-0432.ccr-15-2997.
33. Huang Y -q., Liang C -s. C -h., He L, Tian J, Liang C -s. C -h., Chen X, et al. (2016) Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer SUPPLEMENT. J Clin Oncol:JCO659128. https://doi.org/10.1200/jco.2015.65.9128.
34. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. (2015) CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 114:345–50. https://doi.org/10.1016/j.radonc.2015.02.015.

35. Maas M, Lambregts DM, Nelemans PJ, Heijnen LA, Martens MH, Leijtens JW, et al. (2015) Assessment of Clinical Complete Response After Chemoradiation for Rectal Cancer with Digital Rectal Examination, Endoscopy, and MRI: Selection for Organ-Saving Treatment. Ann Surg Oncol 22:3873–80. https://doi.org/10.1245/s10434-015-4687-9.