Radiology

Predicting Rectal Cancer Response to Neoadjuvant Chemoradiotherapy Using Deep Learning of Diffusion Kurtosis MRI

*Xiao-Yan Zhang, PhD** • *Lin Wang, MD** • *Hai-Tao Zhu, PhD* • *Zhong-Wu Li, MD* • *Meng Ye, BD* • *Xiao-Ting Li, MM* • *Yan-Jie Shi, MD* • *Hui-Ci Zhu, MD* • *Ying-Shi Sun, MD*

From the Departments of Radiology (X.Y.Z., H.T.Z., M.Y., X.T.L., Y.J.S., H.C.Z., Y.S.S.), Gastrointestinal Surgery (L.W.), and Pathology (Z.W.L.), Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital & Institute, No. 52 Fu Cheng Rd, Hai Dian District, Beijing 100142, China. Received April 30, 2019; revision requested July 3; revision received January 16, 2020; accepted January 28. **Address correspondence to** Y.S.S. (e-mail: *sys27@163.com*).

Supported by the National Natural Science Foundation of China (81971584, 91959116), Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (ZYLX201803), Beijing Hospitals Authority Youth Programme (QML20181103), Beijing Hospitals Authority Ascent Plan (20191103), National Key Research and Development Program of China (2019YFC0117705, 2017YFC1309101, 2017YFC1309104), National Science and Technology Major Project (2020ZX09201023), and Beijing Municipal Science and Technology Commission (Z171100001017102).

* X.Y.Z. and L.W. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the editorial by Koh in this issue.

Radiology 2020; 296:56–64 • https://doi.org/10.1148/radiol.2020190936 • Content codes: GI IN

Background: Preoperative response evaluation with neoadjuvant chemoradiotherapy remains a challenge in the setting of locally advanced rectal cancer. Recently, deep learning (DL) has been widely used in tumor diagnosis and treatment and has produced exciting results.

Purpose: To develop and validate a DL method to predict response of rectal cancer to neoadjuvant therapy based on diffusion kurtosis and T2-weighted MRI.

Materials and Methods: In this prospective study, participants with locally advanced rectal adenocarcinoma (≥cT3 or N+) proved at histopathology and baseline MRI who were scheduled to undergo preoperative chemoradiotherapy were enrolled from October 2015 to December 2017 and were chronologically divided into 308 training samples and 104 test samples. DL models were constructed primarily to predict pathologic complete response (pCR) and secondarily to assess tumor regression grade (TRG) (TRG0 and TRG1 vs TRG2 and TRG3) and T downstaging. Other analysis included comparisons of diffusion kurtosis MRI parameters and subjective evaluation by radiologists.

Results: A total of 383 participants (mean age, 57 years ± 10 [standard deviation]; 229 men) were evaluated (290 in the training cohort, 93 in the test cohort). The area under the receiver operating characteristic curve (AUC) was 0.99 for the pCR model in the test cohort, which was higher than the AUC for raters 1 and 2 (0.66 and 0.72, respectively; $P < .001$ for both). AUC for the DL model was 0.70 for TRG and 0.79 for T downstaging. AUC for pCR with the DL model was better than AUC for the best-performing diffusion kurtosis MRI parameters alone (diffusion coefficient in normal diffusion after correcting the non-Gaussian effect [D_{app} value] before neoadjuvant therapy, AUC = 0.76). Subjective evaluation by radiologists yielded a higher error rate (1 – accuracy) (25 of 93 [26.9%] and 23 of 93 [24.8%] for raters 1 and 2, respectively) in predicting pCR than did evaluation with the DL model (two of 93 [2.2%]); the radiologists achieved a lower error rate (12 of 93 [12.9%] and 13 of 93 [14.0%] for raters 1 and 2, respectively) when assisted by the DL model.

Conclusion: A deep learning model based on diffusion kurtosis MRI showed good performance for predicting pathologic complete response and aided the radiologist in assessing response of locally advanced rectal cancer after neoadjuvant chemoradiotherapy.

Online supplemental material is available for this article.

Neoadjuvant chemoradiotherapy (NCRT) has been proven effective in the downstaging of locally advanced rectal cancer, and it leads to pathologic complete response (pCR) in about 20% of patients (1–4). The patient's response to NCRT is particularly important for prognosis and management decisions (5). At present, therapy response is primarily determined by histopathologic assessment after surgery (5), including determination of posttreatment pathologic T and N stage; circumferential resection margin status; and tumor regression grading (TRG) (6). However, this information could be useful in directing the approach to surgery if it were available prior to resection (3,5). Accordingly, development of a noninvasive response evaluation model could aid in the identification of patients with good response who are likely to benefit from local resection, as well as those achieving pCR who might benefit from a watch-and-wait or nonsurgical strategy (2).

The combination of diffusion- and T2-weighted imaging has shown certain advantages in response evaluation (7–10), but to our knowledge, an accurate preoperative therapy response evaluation system has yet to be developed (5). One possible reason is the limitation of a simple diffusion model

Abbreviations

AUC = area under the receiver operating characteristic curve, CI = confidence interval, DKI = diffusion kurtosis imaging, DL = deep learning, NCRT = neoadjuvant chemoradiotherapy, pCR = pathologic complete response, ROI = region of interest, TRG = tumor regression grading

Summary

Deep learning models for diffusion kurtosis MRI predicted pathologic complete response and tumor regression grade and improved subjective evaluation by radiologists.

Key Results

- A deep learning (DL) model for diffusion kurtosis MRI showed excellent performance in predicting pathologic complete response (pCR) of rectal cancer after neoadjuvant chemoradiotherapy and was superior to assessment by two radiologists (area under the receiver operating characteristic curve, 0.99 vs 0.66 and 0.72, respectively).
- Evaluation by radiologists resulted in a higher error rate for predicting pCR when compared with the DL model (26.9% and 24.8% for raters 1 and 2, respectively; 2.2% for the DL model).
- The error rate was reduced when radiologists were assisted by the DL model (13% and 14% for raters 1 and 2, respectively).

in describing the non-Gaussian behavior of water molecular movement in the complex tumor environment (11). Several methods have been proposed to fit the non-Gaussian diffusion curve, including stretched exponential, intravoxel incoherent motion, and diffusion kurtosis imaging (DKI) models (12–14). Among these methods, diffusion kurtosis MRI is relatively robust, as it uses polynomials to fit two unknown parameters (D_{app} and K_{app}) (11,15–17). The parameter K_{app} reflects non-Gaussian diffusion behavior, and D_{app} is the diffusion coefficient in normal diffusion after correcting the non-Gaussian effect. Diffusion kurtosis MRI yields more information on tissue structure than does standard monoexponential diffusion-weighted imaging analysis (15), thus giving radiologists an opportunity to gain further insights into tissue characteristics. Several studies have shown the potential benefit of diffusion kurtosis MRI for therapy response evaluation in the setting of locally advanced rectal cancer (16,17).

In this study, we proposed a method to combine diffusion kurtosis MRI and deep learning (DL) to predict the therapy response of locally advanced rectal cancer after NCRT. The study comprises two parts: The first is a prospective study designed to construct DL models in a training cohort. The second is a prospective study using a test cohort to compare the performances of DL models, DKI mean values, and subjective evaluation by radiologists and to determine the ability of DL models to enhance radiologist performance. This study aimed to provide an accurate and semiautomatic way to predict therapy response after NCRT and a quantitative tool to help inform the decision to perform surgery or follow a nonsurgical approach.

Materials and Methods

Study Participants

This single-center prospective study enrolled consecutive participants with rectal cancer from October 2015 to December 2017. The protocol was approved by the medical ethics committee of Beijing Cancer Hospital. All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional or national research committee and the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Figure 1 shows the inclusion and exclusion criteria. Candidates were participants *(a)* with locally advanced rectal adenocarcinoma proved with histopathology and baseline MRI (≥cT3 or N+) and *(b)* who were scheduled to undergo NCRT in our hospital. All study participants provided written informed consent. All candidates were scheduled for two MRI examinations: pre-NCRT MRI within 1 week before initiation of NCRT and post-NCRT MRI within 1 week before surgery. Participants were excluded if *(a)* they had concurrent malignancy at another site, *(b)* they had undergone treatment before enrollment, *(c)* NCRT was incomplete, *(d)* they did not undergo surgery, *(e)* MRI quality was insufficient for measurements, or *(f)* pathology results were unavailable or the patient had mucinous adenocarcinoma.

MRI Data Acquisition and Diffusion Model

All participants underwent MRI at two time points: within 1 week before initiation of NCRT and within 1 week before surgery; these were defined as pre- and post-NCRT MRI, respectively. All MRI examinations were performed with a 3.0-T MRI scanner (Discovery MR750; GE Healthcare, Milwaukee, Wis) using an eight-channel phased-array body coil in the supine position. To reduce colonic motility, 20 mg of scopolamine butylbromide was injected intramuscularly 30 minutes before MRI. Participants were not asked to undergo bowel preparation before MRI. The MRI protocol included acquisition of axial, coronal, and sagittal T2-weighted images; transverse T1-weighted images; and DKI (Table E1 [online]). The scanning parameters of these protocols are summarized in Table E1 (online).

DKI images were obtained by using single-shot echo-planar imaging with 12 b values (0, 20, 50, 100, 200, 400, 600, 800, 1000, 1200, 1400, and 1600 sec/mm^2).

D_{app} and K_{app} images were obtained as follows:

$$\log[S(b)] = \log[S(0)] - bD_{app} + b^2 K_{app} D_{app}^2 / 6, \quad (1)$$

where $S(b)$ is the signal intensity at nonzero b values and $S(0)$ is the signal intensity at a b value of 0 sec/mm^2. Besides D_{app} and K_{app}, another inputted image is S_{app} = log[S(1000 sec/mm^2)], the logarithmic form of diffusion-weighted signal at a b value of 1000 sec/mm^2.

NCRT Treatment

All participants underwent 22-fraction intensity-modulated radiation therapy (18) that was designed to shorten the treatment course and decrease radiation-related toxicity. Capecitabine (825 mg/m^2 given orally twice per day) was administered concurrently with intensity-modulated radiation therapy. All participants underwent total mesorectal excision surgery within 8–10 weeks after completion of intensity-modulated radiation therapy.

Pathologic Assessment of Response

After total mesorectal excision, surgically resected specimens were fixed in formalin for 23–48 hours, axially sectioned into

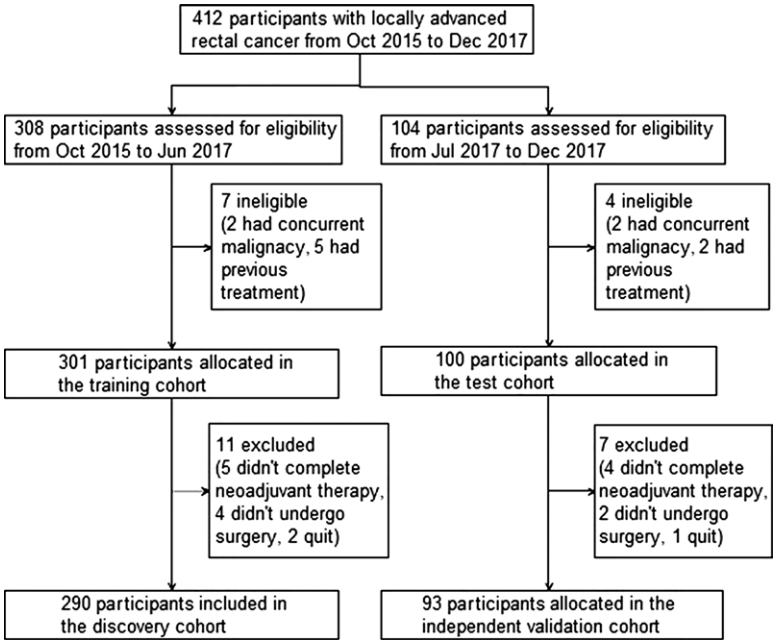

Figure 1: Flowchart shows inclusion and exclusion criteria.

3–5-mm slices, then histopathologically examined and analyzed by two pathologists in consensus (Z.W.L., Y.S.; 10 and 15 years of experience, respectively, in gastrointestinal disease). If there was uncertainty in their evaluation, they discussed the case to reach an agreement. The resected specimens were staged according to the International Union Against Cancer TNM staging system. The absence of any residual cancer cells within the resected surgical specimen was regarded as pathologic complete response (T0N0). TRG was evaluated as described in the National Comprehensive Cancer Network and American Joint Committee on Cancer TRG system. TRG0 and TRG1 were considered good response, whereas TRG2 and TRG3 were considered poor response. T downstage was defined as a pathologic T stage less than T3, and non–T downstage was defined as a pathologic T stage of at least T3. Pathologists evaluated the specimen within 1 week after surgery and were blinded to the results of the radiologic evaluation.

ROI Delineation on MRI Scans

MRI scans were analyzed by two radiologists independently (Z.W.L., Y.S.; 8 and 10 years of experience, respectively, in rectal cancer imaging) within 1 week after the MRI examination. Raters were blinded to the results of histopathology. The regions of interest (ROIs) were created freehand and manually with itk-SNAP software (version 3.8; *www.itksnap.org*) on the pre- and post-NCRT T2-weighted and diffusion-weighted images (*b* value, 1000 sec/mm^2), including the whole tumor and excluding the intestinal lumen. ROIs in the rectal tumor were manually drawn on each image slice that contained the tumor. ROIs were drawn along the contour of the tumor on T2-weighted images (slightly high signal intensity) and contained the surrounding tumor chords and burrs that extended into the mesorectum. ROIs also were placed on the tumor region (high-signal-intensity region, compared with the adjacent normal rectal wall) on diffusion-weighted images (*b* value, 1000 sec/mm^2). Two examples of ROI creation on diffusion-weighted images are shown in Figure E1 (online), with reproducibility evaluation described in Appendix E1 (online).

DL and Model Construction

Model construction.—There was a training data set for model construction and internal validation and a test data set for external validation. We intended to establish three models (Fig E2 [online]) with the same data sets. We set non-pCR and pCR as the labels for model A; bad response (TRG2 and TRG3) and good response (TRG0 and TRG1) as the labels for model B; and non–T downstage and T downstage as the labels for model C. Models A, B, and C all used the same network architecture. The code has been uploaded to GitHub *(https://github.com/radiologypkucancer/rectal_MR_DL/)*. It includes all the procedures used for data preprocessing, model training, and validation.

Data preprocessing.—Data preprocessing included patching, resizing, normalization, and augmentation, which are described in Appendix E1 (online).

Network architecture.—A multipath convolutional neural network with eight inputs was designed to include pre- and post-T2-weighted imaging, pre-D_{app}, post-D_{app}, pre-K_{app}, post-K_{app}, pre-S_{app}, and post-S_{app} data, as shown in Figure E3 (online) (19). Each data stream was processed independently. Thus, useful features could be extracted from different input data streams. Because each data stream may contain different information, the network lets it pass through a different densely connected multimax pooling convolutional neural network. Figure 2 shows the network architecture. Table E2 (online) gives the detailed parameters of the network.

The multimax pooling layer is densely connected. This makes the network convergence faster because there are more paths for gradient backpropagation (20,21).

Training.—The network architecture was implemented using Python 3.6 software based on the Keras 2.1.5 *(https://github.com/keras-team/keras)* DL library with TensorFlow 1.4.0 (22) as its backend. It was trained on a workstation with one NVIDIA TITAN X GPU. To train the network, we used the stochastic gradient descent algorithm with the adaptive moment estimation algorithm (known as ADAM) optimizer (23) with a learning rate of 7×10^{-6} and a decay rate of 10^{-5}, a mini-batch size of 30, and a binary cross-entropy loss function. All parameters were initialized randomly from a Gaussian distribution using the He et al method (24). We also used batch normalization (25) for the intermediate responses after each convolution layer to accelerate the convergence and reduce overfitting. We used a dropout (26) with 0.5 probability and L2 regularization with a λ_2 of 0.025 penalty on neuron weights. The network was trained for 500 epochs.

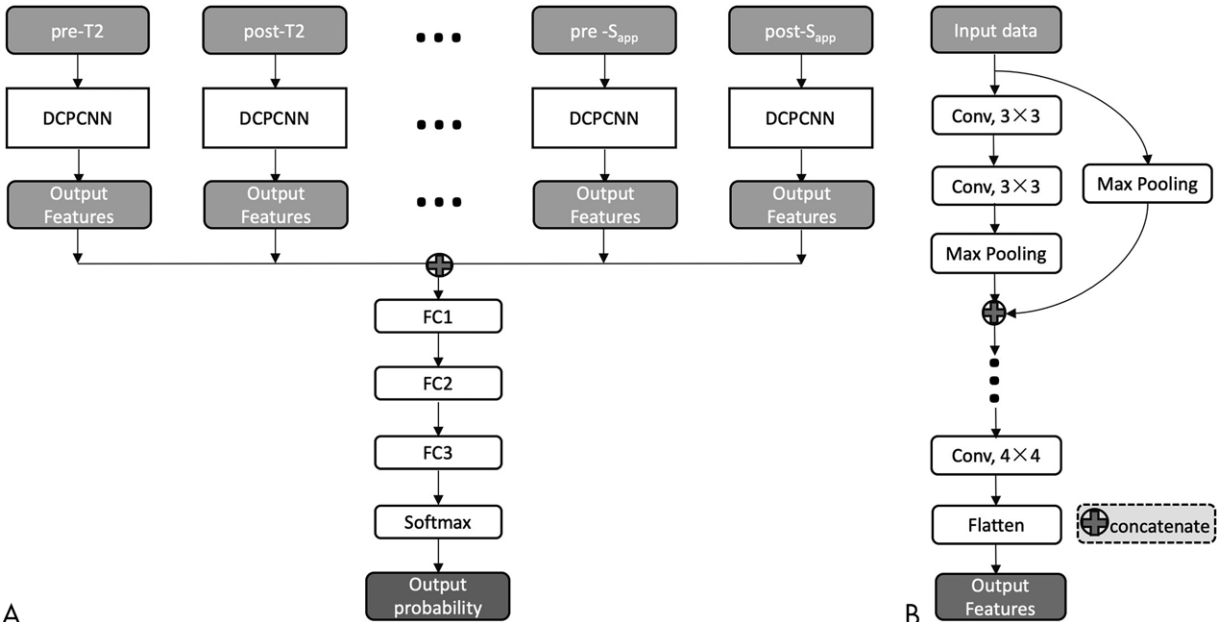

Figure 2: *A*, Multipath deep convolutional neural network architecture that contains, *B*, uniquely designed densely connected multimax pooling convolutional neural network modules used for feature extraction from multicontrast MRI.DCPCNN = densely connected multimax pooling convolutional neural network, FC = fully connected layer.

***Prediction.*—**Given a three-dimensional input data of eight kinds of MRI parameter map, the trained network can predict the two-class probability of aimed label (non-pCR or pCR, TRG0 and TRG1, or TRG2 and TRG3, T downstage or non–T downstage). Classification accuracy of prediction and receiver operating characteristic curve analysis can be conducted based on the two-class prediction probability of each aimed label.

Subjective Evaluation

Raters 1 and 2 were blinded to the pathologic results and independently evaluated all MRI scans of participants in the test cohort within 1 week after the MRI examination. They evaluated tumor stage and lymph node status on both pre- and posttherapy MRI scans and evaluated pCR status, TRG, and T downstage (27,28). Criteria and reproducibility evaluation are detailed in Appendix E1 (online). Subjective evaluation was then repeated by considering the result of the DL models as assistance.

Statistical Analysis

In this study, we used a training cohort to construct the response prediction model and a test cohort chronologically. The primary analysis of this study was to construct a DL model for pCR prediction (model A), which was used to calculate sample size. We arbitrarily supposed that to be useful clinically, a diagnostic test must have an area under the receiver operating curve (AUC) greater than 0.85 (according to the joint determination by clinical experts and methodologists). At least 270 cases (54 participants with pCR, 216 participants with non-pCR) were needed to detect a difference of 0.10 with 85% power and using a two-sided *z* test at a significance level of .05. In consideration of 12% withdraws, we needed to enroll no fewer than 308 participants (62 participants with pCR, 246 participants with non-pCR) to construct the model. The AUC of the constructed model was compared with 0.85 by using a two-sided z test; if a significant result was obtained, analysis was conducted in the test cohort with a sample size of 104 participants for external validation. Thus, eligible participants were continually enrolled in the study until the total number of participants reached the planned sample size of 412. The secondary analysis included constructing two prediction models for TRG (model B) and T downstage (model C) and validating the model in the test cohort.

The AUC of model A was compared with AUCs of models B and C and was also compared with the mean values of DKI parameters by using the DeLong method (29). Bonferroni correction was used for multiple comparisons of AUCs. Other analyses included comparisons of accuracies in pCR prediction among model A, subjective evaluation, and subjective evaluation assisted by model A. *P* values of other analysis were not adjusted for testing multiple hypotheses.

Statistical analysis was conducted with R software (version 3.5.2; *http://www.r-project.org/*). A χ^2 test was used to compare differences in categorical variables. The independent *t* test or Mann-Whitney *U* test was used to compare differences in continuous variables. Sensitivity, specificity, positive and negative predictive values, and accuracy were acquired by selecting the cutoff at the maximum Youden index. The aggregated number of wrong decisions is displayed as error rate (1 – accuracy), false-negative predictive value (1 – sensitivity), and false-positive predictive value (1 – specificity) for model evaluation and subjective evaluation. Decision curve analysis was used to determine the clinical usefulness of this therapy response prediction model (30).

Results

Participant Characteristics

From October 2015 to December 2017, 412 candidates were identified, and 383 met inclusion criteria and were enrolled (229 male participants [59.8%]; mean age, 57 years ± 10 [standard deviation]). The distribution of participant characteristics, including age, sex, histologic grade, pre- and post-NCRT tumor and lymph node stage, TRG, and pCR status, was similar between the training and test cohorts (Table 1).

Table 1: Characteristics of Study Participants in the Training and Test Cohort

Characteristic	Training Cohort (n = 290)	Test Cohort (n = 93)	P Value
Age (y)*	56 ± 10	58 ± 9	.13
Sex	…	…	.38
Male	177 (61.0)	52 (55.9)	…
Female	113 (39.0)	41 (44.1)	…
Histologic grade	…	…	.96
I	4 (1.4)	2 (2.2)	…
II	248 (85.5)	80 (86.0)	…
III	24 (8.3)	7 (7.5)	…
IV	12 (4.1)	3 (3.2)	…
V	2 (0.7)	1 (1.1)	…
Pre-NCRT T stage	…	…	.94
T0	0 (0)	0 (0)	…
T1	0 (0)	0 (0)	…
T2	25 (8.6)	9 (9.7)	…
T3	208 (71.7)	67 (72.0)	…
T4a	33 (11.4)	11 (11.8)	…
T4b	24 (8.3)	6 (6.5)	…
Pre-NCRT N stage	…	…	.31
N0	11 (3.8)	4 (4.3)	…
N1a	10 (3.4)	5 (5.4)	…
N1b	40 (13.8)	6 (6.5)	…
N1c	1 (0.3)	1 (1.1)	…
N2a	82 (28.3)	33 (35.5)	…
N2b	146 (50.3)	44 (47.3)	…
pCR	…	…	>.99
pCR	56 (19.3)	18 (19.4)	…
Non-pCR	234 (80.7)	75 (80.6)	…
TRG	…	…	.89
TRG0	66 (22.8)	18 (19.4)	…
TRG1	99 (34.1)	33 (35.5)	…
TRG2	117 (40.3)	40 (43.0)	…
TRG3	8 (2.8)	2 (2.1)	…
Pathologic T stage	…	…	.80
T0	66 (22.8)	18 (19.4)	…
T1	17 (5.9)	4 (4.3)	…
T2	83 (28.6)	27 (29.0)	…
T3	123 (42.4)	43 (46.2)	…
T4a	1 (0.3)	1 (1.1)	…
Pathologic N stage	…	…	.37
N0	206 (71.0)	71 (76.3)	…
N1a	36 (12.4)	9 (9.7)	…
N1b	23 (7.9)	5 (5.4)	…
N1c	9 (3.1)	4 (4.3)	…
N2a	7 (2.4)	4 (4.3)	…
N2b	9 (3.1)	0 (0)	…

Note.—Unless otherwise indicated, data are number of participants, and data in parentheses are percentages. NCRT = neoadjuvant chemoradiotherapy, pCR = pathologic complete response, TRG = tumor regression grade.

* Data are mean ± standard deviation.

NCRT Therapy Response Prediction Model Construction with DL

In the training cohort, the AUC of model A for pCR prediction was 0.997 (95% confidence interval [CI]: 0.995, 1.000). The AUCs of models B and C were 0.99 (95% CI: 0.98, 1.00) and 0.99 (95% CI: 0.99, 1.00), respectively. In the test cohort, the AUCs of models A, B, and C were 0.99 (95% CI: 0.94, 1.00), 0.70 (95% CI: 0.59, 0.79), and 0.79 (95% CI: 0.69, 0.87), respectively (Fig 3, Table 2).

As shown in Table 2, the performance of model A was better than the performance of models B and C (AUC, 0.99 vs 0.70 and 0.79; $P < .01$ and $P = .03$, respectively). The performance of the mean value of pre-D_{app}, pre-K_{app}, post-D_{app}, post-K_{app}, ΔD_{app}, and ΔK_{app} in the prediction of therapy response is shown in Tables E3–E5 (online). We also found that the performance of model A was better than the mean pre-D_{app} value (the parameter with the largest AUC in Table E3 [online]) for pCR classification (AUC, 0.99 vs 0.76; $P = .01$).The performance of model B was similar to the ΔD_{app} value (the parameter with the largest AUC in Table E4 [online]) in the differentiation of TRG0 and TRG1 from TRG2 and TRG3 (AUC, 0.70 vs 0.63; $P = .45$). The performance of model C was better than the ΔK_{app} value (the parameter with the largest AUC in Table E5 [online]) for T downstage classification (AUC, 0.79 vs 0.64; $P = .03$). The diagnostic accuracy of D_{app} and K_{app} is given in Appendix E1 (online), and the parameter showing the best accuracy was chosen for comparison with the corresponding DL model.

Decision Curve Analysis for DL Model

The decision curve analysis of model A for pCR is presented in Figure E4 (online). The decision curve shows that the DL model adds more benefit than the treat-all-participants scheme or the treat-none scheme in a wide range of threshold probabilities.

Comparison between Subjective Evaluation and Model Prediction

Both raters had worse diagnostic performance than model A in the assessment of pCR (both $P < .01$ for comparison of total accuracy) with sensitivity, specificity, positive and negative predictive values, and total accuracy of 55.6%, 77.3%, 37.1%, 87.9%, and 73.1%, respectively, for rater 1 and 66.7%, 77.3%, 41.4%, 90.6%, and 75.2%, respectively, for rater 2. With the as-

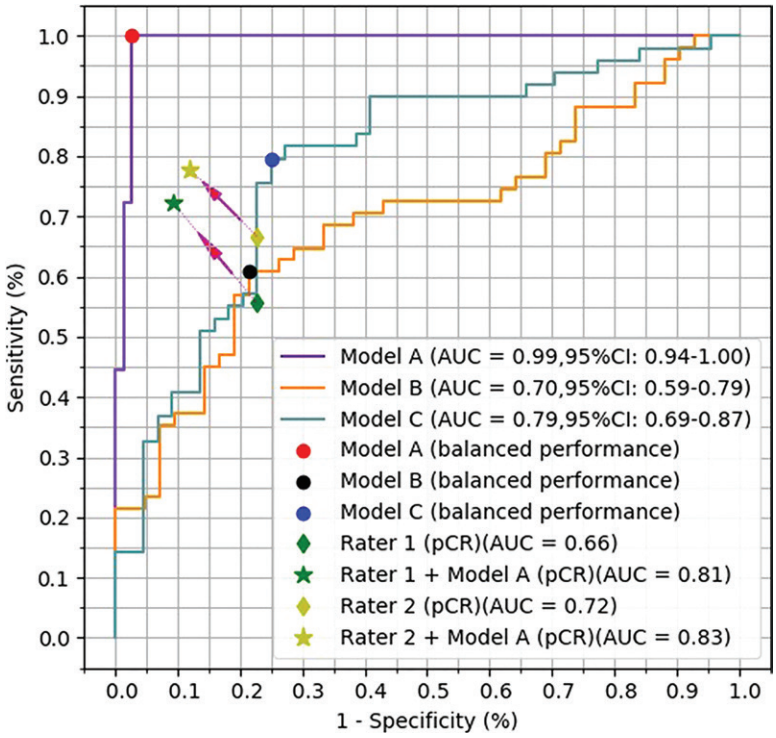

Figure 3: Receiver operating characteristic (ROC) curve analysis of deep learning models for pathologic complete response (pCR) (model A), tumor regression grade (TRG) (model B), and T downstage (model C) prediction shows the diagnostic performance of subjective evaluation (raters 1 and 2) for pCR prediction. Balanced performance is the pair of sensitivity and specificity with the maximum Youden index (sensitivity + specificity - 1). Diagnostic performance improved significantly (arrows) after the assistance of model A (P = .002 for rater 1, P = .01 for rater 2). AUC = area under the ROC curve.

sistance of model A, both raters showed better performance (P = .02 and P = .04 for the comparison of total accuracy), with sensitivity, specificity, positive and negative predictive values, and total accuracy of 72.2%, 90.7%, 65.0%, 93.2% and 87.1%, respectively, for rater 1 and 77.8%, 88.0%, 60.9%, 94.3%, and 86.0%, respectively, for rater 2 (Fig 3). With the assistance of model A, the error rates of raters 1 and 2 were 12.9% and 14.0%, respectively (26.9% and 24.8%, respectively, without the assistance of model A); the false-negative predictive values of raters 1 and 2 were 9.3% and 12%, respectively (22.7% and 22.7%, respectively, without the assistance of model A); the false-positive predictive values of raters 1 and 2 were 27.8% and 26.3%, respectively (44.4% and 33.3%, respectively, without the assistance of model A), and were comparable to those obtained with model A (Fig 4). The subjective evaluation results for T downstage and TRG are listed in Table E6 (online). Figure E5 (online) shows the comparison of sensitivity, specificity, positive and negative predictive values, and total accuracy among subjective evaluation, model evaluation, and subjective evaluation with the assistance of model A for pCR prediction.

The AUCs of raters in assessing pCR were 0.66 (95% CI: 0.56, 0.76) for rater 1 and 0.72 (95% CI: 0.62, 0.81) for rater 2, which were significantly inferior to the AUC of model A (both $P < .001$) (Fig 3). With the assistance of model A, both raters obtained significantly higher AUCs (0.82 [95% CI: 0.72, 0.89] for radiologist 1 [P = .002] and 0.83 [95% CI: 0.74, 0.90] for radiologist 2 [P = .01]) (Fig 3).

Discussion

The ability to predict pathologic complete response (pCR), tumor regression grading (TRG), and T downstaging at preoperative imaging after neoadjuvant therapy may improve outcomes for patients by appropriately allocating them to undergo surgical resection or nonsurgical treatment. We aimed to evaluate the ability of a deep learning (DL) model incorporating diffusion kurtosis imaging (DKI) and T2-weighted images to classify pCR, TRG, and T downstaging. The DL model for pCR prediction showed good diagnostic performance, with an area under the receiver operating curve (AUC) of 0.99 (95% confidence interval [CI]: 0.94, 1.00), which was better than the best predictor using the DKI mean value (AUC, 0.76; 95% CI: 0.71, 0.80; P = .01). However, the DL model for TRG or T downstaging prediction showed relatively poorer accuracy, which was similar with the DKI mean value.

In addition, this study found that with the assistance of the DL model, radiologists had lower error rates for subjective pCR prediction (error rate of rater 1 was reduced from 26.9% to 12.9%, error rate of radiologist 2 was reduced from 24.8% to 14.0%), which brought the error rate in line with error rates seen with the DL model alone. These results suggest that the application of artificial intelligence in assisting with clinical diagnosis may reduce the error rate for pCR prediction and could improve confidence in management decisions based on preoperative MRI findings. The performance of models B and C was not as good as the performance of model A, but it does not mean these two DL networks failed. The accuracy of the DL model relies on both the structure of the network and the label of ground truth. It is known that pathologic evaluation of TRG and T downstaging are more subjective than pCR evaluation (31,32). The lack of a precise and objective pathologic label for TRG and T downstaging could be a reason for the poor performance of the DL models. We also noticed that radiologists' accuracies decreased when using DL assistance to evaluate TRG and T downstaging. The difference in the performance of DL assistance may also come from radiologists' confidence in the DL model. If the DL model gives opposite predictions for some easy cases that the radiologists believe in themselves, they may be skeptical of the DL results. In this situation, even DL is slightly more accurate than radiologists, but the radiologists may not believe the DL results, thereby causing a decline in accuracy after DL assistance (Fig 5; Figs E6, E7 [online]).

Standard diffusion-weighted imaging generally scans two images, one at a high b value, such as 1000 sec/mm^2, and another at a b value of 0 sec/mm^2. Standard apparent diffusion coefficients are calculated based on the monoexponential decay depicted by a Gaussian model. On the contrary, the DKI model considers the non-Gaussian diffusion effect, where D_{app} is the linear term and K_{app} is the quadratic term. K_{app} is a unique value in DKI that has

Table 2: Model Performance of Therapy Response Prediction for Participants with Locally Advanced Rectal Cancer in the Test Cohort

Model	AUC	Sensitivity	Specificity	PPV	NPV	Total Accuracy
A	0.99 [0.94, 1.00]	18/18 (100) [100, 100]	73/75 (97.3) [93.7, 100]	18/20 (90.0) [76.5, 100]	73/73 (100) [100, 100]	91/93 (97.8) [95.4, 100.0]
B	0.70 [0.59, 0.79]	11/18 (60.8) [47.4, 74.2]	59 (78.6) [66.2, 90.9]	11/27 (40.7) [17.6, 63.8]	59/66 (89.4) [80.2, 98.6]	70/93 (75.3) [62.4, 88.2]
C	0.79 [0.69, 0.87]	14/18 (77.8) [67.3, 88.9]	56/75 (74.7) [62.2, 87.8]	14/33 (42.4) [20.5, 63.3]	56/60 (93.3) [84.1, 100]	70/93 (75.3) [62.4, 88.2]

Note.—Unless otherwise indicated, data are proportions, data in parentheses are percentages, and data in brackets are 95% confidence intervals. Model A is pathologic complete response prediction model; model B, tumor regression grade prediction model; and model C, T downstage model. AUC = area under the receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value.

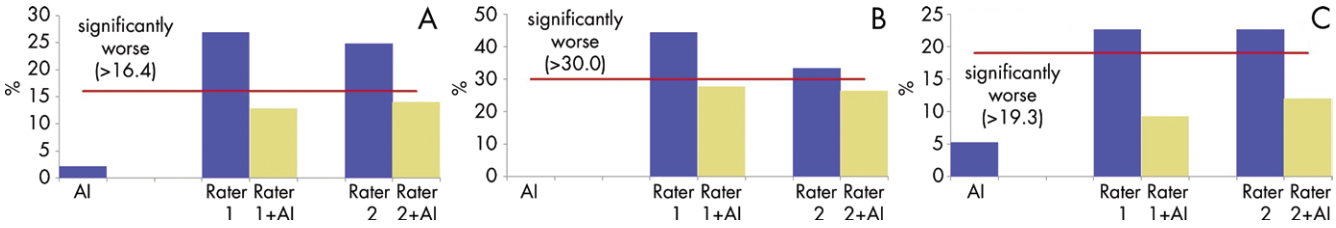

Figure 4: Bar graphs show comparison of, *A*, error rate, *B*, false-negative predictive (FNP) value, and, *C*, false-positive predictive (FPP) value among subjective evaluation, model evaluation, and subjective evaluation with the assistance of model A for pathologic complete response (pCR) prediction. A lower error rate and lower FNP and FPP values indicate better diagnostic performance. The error rate and FNP and FPP values of subjective evaluation for pCR prediction was higher than for model A for raters 1 and 2; with the assistance of model A, subjective evaluation had a similar error rate and FNP and FPP values as model A alone for raters 1 and 2. Error rate was defined as the aggregated number of wrong decisions (100% – accuracy), which was two of 93 (2.2%), 25 of 93 (26.9%), 12 of 93 (12.9%), 23 of 93 (24.8%), and 13 of 93 (14%) for artificial intelligence (AI), rater 1, rater 1 and AI, rater 2, and rater 2 and AI, respectively. FNP value was defined as 100% minus sensitivity and was 0 of 18 (0%), eight of 18 (44.4%), five of 18 (27.8%), six of 18 (33.3%), and four of 18 (22.2%) for AI, rater 1, rater 1 and AI, rater 2, and rater 2 and AI, respectively. FPP value was defined as 100% minus specificity and was two of 75 (2.7%), 17 of 75 (22.7%), seven of 75 (9.3%), 17 of 75 (22.7%), and nine of 75 (12.0%) for AI, rater 1, rater 1 and AI, rater 2, and rater 2 and AI, respectively.

no counterpart in conventional diffusion-weighted imaging. The fitting of a DKI model requires at least three *b* values. Generally, apparent diffusion coefficients can be replaced by D_{app} because they have the same unit of measure (seconds per square millimeter). D_{app} is more accurate than the apparent diffusion coefficient for two reasons: First, by removing the non-Gaussian effect, D_{app} may better quantify normal Gaussian diffusion. Second, since D_{app} is calculated with multiple *b* values, it is insensitive to the artifacts or abnormal noise on any one image. K_{app} as the non-Gaussian component may depict the inhomogeneity of diffusion that cannot be measured with conventional diffusion-weighted imaging. Our research attempts to explore the potential of DKI in the therapy response evaluation of rectal cancer. The results suggest that DKI is a promising tool in clinical rectal radiology but that it requires further investigation.

To our knowledge, this is the first application of a DKI and T2-weighted imaging DL model in therapy response evaluation in participants with locally advanced rectal cancer after NCRT. The decision curve (Fig E4 [online]) shows that the net benefit of DL model A is more than that of the treat-all-participants scheme or the treat-none scheme at a wide range of threshold probabilities. Participants with a pCR represent a distinct subgroup of good responders. If participants with higher pCR probability can be identified before surgery, the necessity of total mesorectal excision surgery for these participants could be reevaluated because the long-term survival rate of patients who had pCR after rectum resection is similar to that of the watch-and-wait group, in which the local regrowth rate is only 24%–25%, and the salvage resection rate is higher than 85% (2,33,34).

Our study had some limitations. The main limitation of this single-center study was the lack of a true external validation cohort in which different MRI scanners and field strengths were used. A well-validated DL model based on the data of one center may not work at another center with a different scanner or imaging configuration. Therefore, multicenter investigation is necessary to generalize the DL model from bench to bedside. Secondly, manual delineation of the ROI may introduce subjectivity and variability. In this work, good interobserver consistency is obtained by two experienced radiologists with a Dice similarity coefficient larger than 0.8. DL models trained with an ROI delineated by two raters shows a similar result (AUC = 0.99 for pCR prediction by both sets of ROIs). However, interobserver consistency between inexperienced radiologists has not been analyzed, and the effect of inconsistent ROI placement on final DL model performance was not evaluated in our current study. Although automatic segmentation makes it possible to learn delineation from experienced radiologists, the accuracy of automatic segmentation is still a challenge for rectal cancer, especially after NCRT (35). Third, DKI has not been used as widely in clinical practice as diffusion-weighted imaging due to the need for a higher *b* value and more *b* values (longer scanning time). At

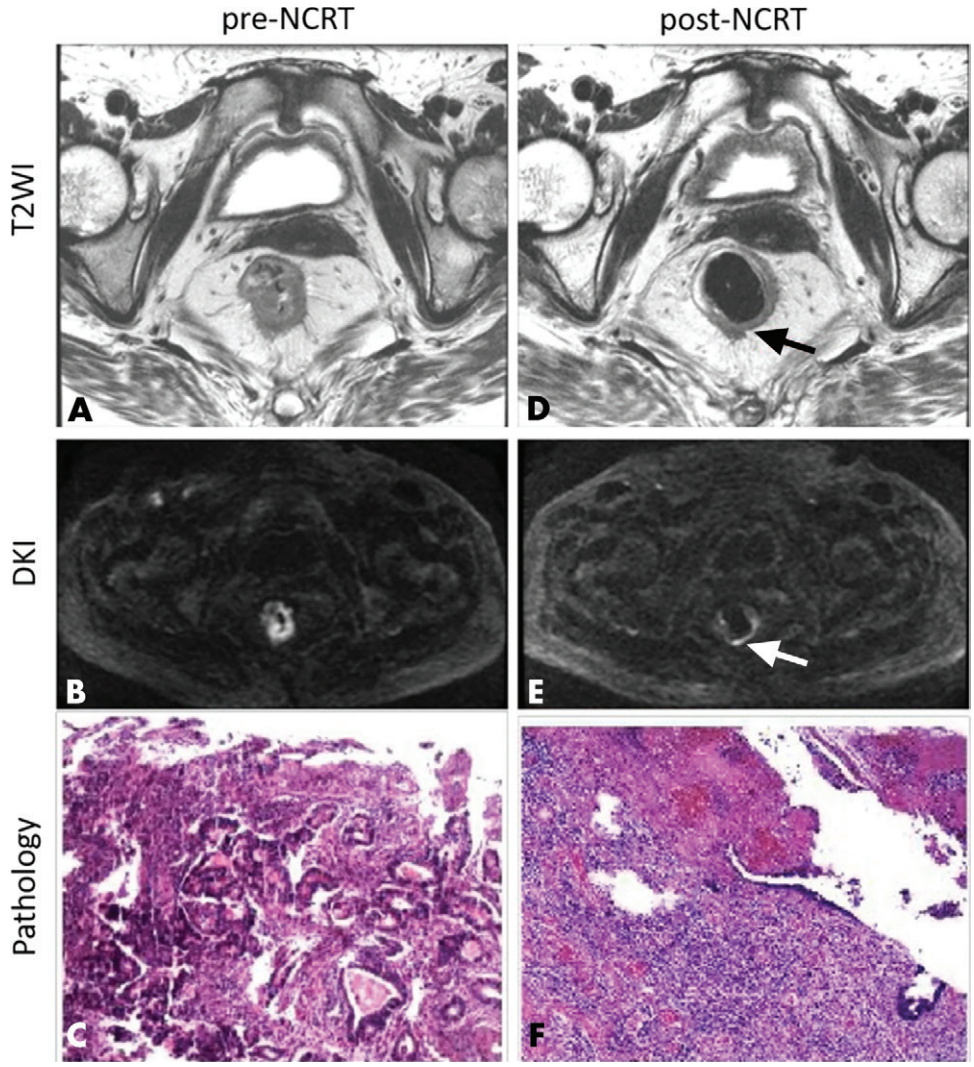

Figure 5: Pre- and post-neoadjuvant chemoradiotherapy (NCRT), T2-weighted (T2WI), diffusion kurtosis (DKI), and pathologic images in a 62-year-old woman with rectal adenocarcinoma. Pre-NCRT, *A*, T2-weighted and, *B*, diffusion kurtosis images show tumor location at the 1–9 o'clock position within the rectal wall. *C*, Colonoscopy biopsy specimen shows a rectal adenocarcinoma. (Hemotoxylin-eosin staining; original magnification, ×100.) Post-NCRT, *D*, T2-weighted (arrow) and, *E*, diffusion kurtosis images show obvious tumor regression (arrow) compared with, *A*, and, *B*, only tumor residual located at the 3–6 o'clock position within the rectal wall. Radiologists' subjective evaluation is non–pathologic complete response (pCR), even though the deep learning model suggested a diagnosis of pCR. *F*, At pathologic analysis after surgery, this patient was shown to have pCR. (Hemotoxylin-eosin staining; original magnification, ×100.)

present, the application of DKI technology favors nervous system imaging, and only a limited number of DKI studies on rectal tumors have been reported (16,17). To our knowledge, there is no agreed-upon standard *b* value setting for body DKI imaging. Finally, it is difficult to quantify the relative contributions of T2-weighted imaging and DKI to the final model accuracy due to the nonlinear calculation in the final layers of the network. If each component (T2-weighted imaging, D_{app}, K_{app}, and S_{app}) is used as the only channel to predict pCR, DKI components produce higher accuracy than the T2-weighted imaging component. From this point of view, the contribution of DKI is larger than that of T2-weighted imaging. However, T2-weighted imaging is indispensable in clinical practice for radiologists to evaluate rectal cancer. It is reasonable to include T2-weighted imaging in the DL model.

In conclusion, based on pre- and post-neoadjuvant chemoradiotherapy diffusion kurtosis MRI, we developed a pathologic complete response (pCR) prediction model by using deep learning and obtained excellent performance in our test cohort. Additionally, the performance of the radiologist's subjective evaluation of pCR was improved with the assistance of the model, suggesting that the model may provide an effective diagnostic reference for pCR evaluation in clinical routine use.

Author contributions: Guarantors of integrity of entire study, X.Y.Z., L.W., Z.W.L., H.C.Z., Y.S.S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, X.Y.Z., H.T.Z., Z.W.L., X.T.L.,

H.C.Z., Y.S.S.; clinical studies, X.Y.Z., L.W., Z.W.L., Y.J.S., H.C.Z., Y.S.S.; statistical analysis, X.Y.Z., H.T.Z., Z.W.L., X.T.L., H.C.Z., Y.S.S.; and manuscript editing, X.Y.Z., H.T.Z., Z.W.L., X.T.L., Y.J.S., H.C.Z., Y.S.S.

Disclosures of Conflicts of Interest: **X.Y.Z.** disclosed no relevant relationships. **L.W.** disclosed no relevant relationships. **H.T.Z.** disclosed no relevant relationships. **Z.W.L.** disclosed no relevant relationships. **M.Y.** disclosed no relevant relationships. **X.T.L.** disclosed no relevant relationships. **Y.J.S.** disclosed no relevant relationships. **H.C.Z.** disclosed no relevant relationships. **Y.S.S.** disclosed no relevant relationships.

References

1. Benson AB 3rd, Venook AP, Al-Hawary MM, et al. Rectal Cancer, Version 2.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018;16(7):874–901.
2. van der Valk MJM, Hilling DE, Bastiaannet E, et al. 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 2018;391(10139):2537–2545.
3. Maas M, Nelemans PJ, Valentini V, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. Lancet Oncol 2010;11(9):835–844.
4. Al-Sukhni E, Attwood K, Mattson DM, Gabriel E, Nurkin SJ. Predictors of pathologic complete response following neoadjuvant chemoradiotherapy for rectal cancer. Ann Surg Oncol 2016;23(4):1177–1186.
5. Patel UB, Taylor F, Blomqvist L, et al. Magnetic resonance imaging-detected tumor response for locally advanced rectal cancer predicts survival outcomes: MERCURY experience. J Clin Oncol 2011;29(28):3753–3760.
6. Vecchio FM, Valentini V, Minsky BD, et al. The relationship of pathologic tumor regression grade (TRG) and outcomes after preoperative therapy in rectal cancer. Int J Radiat Oncol Biol Phys 2005;62(3):752–760.
7. Kim SH, Lee JM, Hong SH, et al. Locally advanced rectal cancer: added value of diffusion-weighted MR imaging in the evaluation of tumor response to neoadjuvant chemo- and radiation therapy. Radiology 2009;253(1):116–125.
8. Jacobs L, Intven M, van Lelyveld N, et al. Diffusion-weighted MRI for early prediction of treatment response on preoperative chemoradiotherapy for patients with locally advanced rectal cancer: A feasibility study. Ann Surg 2016;263(3):522–528.
9. Prasad DS, Scott N, Hyland R, Guthrie JA, Tolan DJ. Diffusion-weighted MR imaging for early detection of tumor histopathologic downstaging in rectal carcinoma after chemotherapy and radiation therapy. Radiology 2010;256(2):671–672; author reply 672.
10. Sun YS, Zhang XP, Tang L, et al. Locally advanced rectal carcinoma treated with preoperative chemotherapy and radiation therapy: preliminary analysis of diffusion-weighted MR imaging for early detection of tumor histopathologic downstaging. Radiology 2010;254(1):170–178.
11. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology 1988;168(2):497–505.
12. Bennett KM, Schmainda KM, Bennett RT, Rowe DB, Lu H, Hyde JS. Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model. Magn Reson Med 2003;50(4):727–734.
13. Le Bihan D. Apparent diffusion coefficient and beyond: what diffusion MR imaging can tell us about tissue structure. Radiology 2013;268(2):318–322.
14. Iima M, Yano K, Kataoka M, et al. Quantitative non-Gaussian diffusion and intravoxel incoherent motion magnetic resonance imaging: differentiation of malignant and benign breast lesions. Invest Radiol 2015;50(4):205–211.
15. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed 2010;23(7):698–710.
16. Zhu L, Pan Z, Ma Q, et al. Diffusion Kurtosis Imaging Study of Rectal Adenocarcinoma Associated with Histopathologic Prognostic Factors: Preliminary Findings. Radiology 2017;284(1):66–76.
17. Yu J, Xu Q, Song JC, et al. The value of diffusion kurtosis magnetic resonance imaging for assessing treatment response of neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Eur Radiol 2017;27(5):1848–1857.
18. Wheeler JM, Warren BF, Mortensen NJ, et al. Quantification of histologic regression of rectal cancer after irradiation: a proposal for a modified staging system. Dis Colon Rectum 2002;45(8):1051–1056.
19. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Eprint arXiv:1409.1556. https://arxiv.org/abs/1409.1556. Published 2014. Accessed April 6, 2020.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; 770–778.
21. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. arXiv:1608.06993. https://arxiv.org/abs/1608.06993. Published 2016. Accessed April 6, 2020.
22. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://arxiv.org/abs/1605.08695. Published 2016. Accessed April 6, 2020.
23. Diederik K, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980. Published 2014. Accessed April 6, 2020.
24. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proceedings of the IEEE International Conference on Computer Vision, 2015; 1026–1034.
25. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167. https://arxiv.org/abs/1502.03167. Published 2015. Accessed April 6, 2020.
26. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.
27. Beets-Tan RG, Beets GL. Rectal cancer: review with emphasis on MR imaging. Radiology 2004;232(2):335–346.
28. Smith JJ, Chow OS, Gollub MJ, et al. Organ Preservation in Rectal Adenocarcinoma: a phase II randomized controlled trial evaluating 3-year disease-free survival in patients with locally advanced rectal cancer treated with chemoradiation plus induction or consolidation chemotherapy, and total mesorectal excision or nonoperative management. BMC Cancer 2015;15(1):767.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
30. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565–574.
31. Smith FM, Wiland H, Mace A, Pai RK, Kalady MF. Clinical criteria underestimate complete pathological response in rectal cancer treated with neoadjuvant chemoradiotherapy. Dis Colon Rectum 2014;57(3):311–315.
32. N Kalimuthu S, Serra S, Dhani N, et al. Regression grading in neoadjuvant treated pancreatic cancer: an interobserver study. J Clin Pathol 2017;70(3):237–243.
33. Smith JJ, Strombom P, Chow OS, et al. Assessment of a watch-and-wait strategy for rectal cancer in patients with a complete response after neoadjuvant therapy. JAMA Oncol 2019;5(4):e185896.
34. Dattani M, Heald RJ, Goussous G, et al. Oncological and survival outcomes in watch and wait patients with a clinical complete response after neoadjuvant chemoradiotherapy for rectal cancer: A systematic review and pooled analysis. Ann Surg 2018;268(6):955–967.
35. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. Sci Rep 2017;7(1):5301 [Published correction appears in Sci Rep 2018;8(1):2589.].