Radiology

ORIGINAL RESEARCH • GASTROINTESTINAL IMAGING

An MRI Deep Learning Model Predicts Outcome in Rectal Cancer

*Xiaofeng Jiang, MMed** • *Hengyu Zhao, MMed** • *Oliver Lester Saldanha, MSc* • *Sven Nebelung, MD, PhD* • *Christiane Kuhl, MD, PhD* • *Iakovos Amygdalos, MD, PhD* • *Sven Arke Lang, MD* • *Xiaojian Wu, MD* • *Xiaochun Meng, MD* • *Daniel Truhn, MSc, MD* • *Jakob Nikolas Kather, MD, MSc*** • *Jia Ke, MD***

From the Departments of Colorectal Surgery and General Surgey (X.J., H.Z., X.W., J.K.) and Radiology (X.M.), the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China; Guangdong Institute of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Guangzhou, China (X.J., H.Z., X.W., X.M., J.K.); Department of Medicine III (X.J., O.L.S., J.N.K.), Department of Diagnostic and Interventional Radiology (X.J., S.N., C.K., D.T.), and Department of Surgery and Transplantation (I.A., S.A.L.), University Hospital RWTH Aachen, Aachen, Germany; and Else Kröner Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany (X.J., O.L.S., J.N.K.). Received September 25, 2022; revision requested December 5; revision received March 29, 2023; accepted April 7. **Address correspondence to** J.N.K. (email: *jakob-nikolas.kather@alumni.dkfz.de*).

X.J. is supported by the program of the China Scholarship Council (no. 202106380048). J.N.K. is supported by the Bundesministerium für Gesundheit (German Federal Ministry of Health) (DEEP LIVER, ZMVI1-2520DAT111), the Max-Eder-Programme of Deutsche Krebshilfe (German Cancer Aid) (grant no. 70113864), the Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research) (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; and TANGERINE, 01KT2302 through ERA-NET Transcan), the Deutscher Akademischer Austauschdienst (German Academic Exchange Service) (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), and the National Institute for Health and Care Research (NIHR213331) at Leeds Biomedical Research Centre. This study was supported by National Key Clinical Discipline and the program of Guangdong Gastrointestinal Disease Research Center (2020B1111170004).

* X.J. and H.Z. contributed equally to this work.

** J.N.K. and J.K. are co–senior authors.

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

See also the editorial by Langs in this issue.

Radiology 2023; 307(5):e222223 • https://doi.org/10.1148/radiol.222223 • Content codes: GI AI MR

Background: Deep learning (DL) models can potentially improve prognostication of rectal cancer but have not been systematically assessed.

Purpose: To develop and validate an MRI DL model for predicting survival in patients with rectal cancer based on segmented tumor volumes from pretreatment T2-weighted MRI scans.

Materials and Methods: DL models were trained and validated on retrospectively collected MRI scans of patients with rectal cancer diagnosed between August 2003 and April 2021 at two centers. Patients were excluded from the study if there were concurrent malignant neoplasms, prior anticancer treatment, incomplete course of neoadjuvant therapy, or no radical surgery performed. The Harrell C-index was used to determine the best model, which was applied to internal and external test sets. Patients were stratified into high- and low-risk groups based on a fixed cutoff calculated in the training set. A multimodal model was also assessed, which used DL model–computed risk score and pretreatment carcinoembryonic antigen level as input.

Results: The training set included 507 patients (median age, 56 years [IQR, 46–64 years]; 355 men). In the validation set (n = 218; median age, 55 years [IQR, 47–63 years]; 144 men), the best algorithm reached a C-index of 0.82 for overall survival. The best model reached hazard ratios of 3.0 (95% CI: 1.0, 9.0) in the high-risk group in the internal test set (n = 112; median age, 60 years [IQR, 52–70 years]; 76 men) and 2.3 (95% CI: 1.0, 5.4) in the external test set (n = 58; median age, 57 years [IQR, 50–67 years]; 38 men). The multimodal model further improved the performance, with a C-index of 0.86 and 0.67 for the validation and external test set, respectively.

Conclusion: A DL model based on preoperative MRI was able to predict survival of patients with rectal cancer. The model could be used as a preoperative risk stratification tool.

Published under a CC BY 4.0 license.

Supplemental material is available for this article.

The current standard of treatment for patients with locally advanced rectal cancer is neoadjuvant therapy followed by total mesorectal excision (1). Although this multimodal treatment results in a lower local recurrence rate (2), it has not significantly improved long-term survival (3–5). The existing stratification of prognosis is mainly based on the established TNM tumor staging system and pathologic tumor regression grading. However, TNM is not an optimal risk assessment tool because of the clinical heterogeneity of patients at the same stage. Tumor regression grading, on the other hand, is only available in postoperative specimens, which limits clinical decision-making at the initial treatment phases (6). Therefore, a model that provides prognostic information at the time of cancer diagnosis would be useful for optimizing treatment and monitoring.

Pelvic MRI is routinely used clinically for the diagnosis and staging of rectal cancer and can be used to detect potential prognostic factors, such as extramural vascular invasion and invasion of the circumferential resection margin (7,8). With the development of

This copy is for personal use only. To order copies, contact *reprints@rsna.org*

Abbreviations

AUC = area under the receiver operating characteristic curve, CEA = carcinoembryonic antigen, CNN = convolutional neural network, DFS = disease-free survival, DL = deep learning, HR = hazard ratio, OS = overall survival, ViT = Vision Transformer

Summary

A Vision Transformer–based deep learning network using preoperative MRI scans as input accurately predicted overall survival and disease-free survival of patients with rectal cancer.

Key Results

- Deep learning networks were trained and tested using pelvic MRI scans from 725 patients with rectal cancer from China, and the Vision Transformer (ViT) outperformed traditional neural networks (overall survival [OS] C-index, 0.82 vs 0.67 [ResNet]; $P = .02$).
- The ViT-computed risk score was associated with 1-, 3-, and 5-year OS in patients (area under the receiver operating characteristic curve, 0.65, 0.64, and 0.74, respectively) in the internal test set.
- The ViT-computed risk score and pretreatment carcinoembryonic antigen level in combination showed better prognostic prediction compared with either variable alone (OS C-index, 0.86 vs 0.82 and 0.61, respectively) in the validation set; the performance was upheld in an external test set obtained from an institution in Germany.

quantitative image analysis methods in the past decade, it is possible to automatically extract quantitative features from medical radiology images by computer (9). Many studies have used classic handcrafted radiomics on MRI scans to predict chemotherapy efficacy, prognosis, and molecular typing among patients with locally advanced rectal cancer (10–12). However, radiomics requires laborious and time-consuming precise tumor labeling, hampering its clinical application; in addition, the preselection of morphologic features constrains the information that can be extracted from the images.

Deep learning (DL), an emerging machine learning technique, is promising for overcoming these problems. In the field of computer vision, Vision Transformer (ViT), a transformer-based model pretrained on 21 000 natural images, has achieved state-of-the-art performance on many tasks (13). Unlike traditional machine learning techniques, DL does not require manual feature extraction and selection by domain experts. It can directly process raw data and automatically develop its own representations needed for pattern recognition, so there is no need to explicitly specify rules or features. When given enough data points, DL has been shown to lead to superior performance compared with radiomic analysis in MRI data in breast cancer (14). Recently, DL techniques have even achieved expert-level performance in various medical image analysis tasks (15,16). Our previous studies have also shown that DL can predict pathologic complete response at pre- and posttreatment MRI (17). However, to our knowledge, there is currently no validated DL method for prognostication in patients with rectal cancer, and some of the more advanced technologies, such as ViT, are not being used.

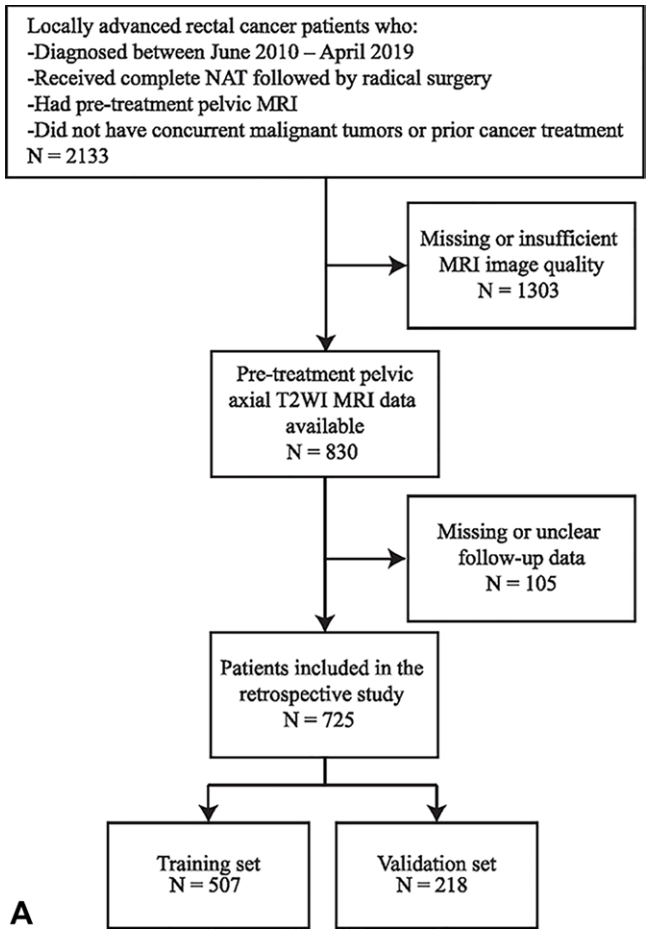

Figure 1: Flowcharts show patient inclusion for the groups with **(A)** locally advanced rectal cancer, which composed the training and validation sets *(Fig 1 continues)*,

In this study, we developed and validated a DL model–computed risk score to predict survival in patients with rectal cancer based on segmented tumor volumes from pretreatment T2-weighted images. In addition, we investigated if the efficacy of the model could be further enhanced by integrating a blood-based tumor biologic marker.

Materials and Methods

Ethics Statement

This retrospective study was performed in accordance with the Declaration of Helsinki and adheres to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis, or TRIPOD, statement (Appendix S1) (18). We analyzed Chinese patients with rectal cancer from the Sixth Affiliated Hospital, Sun Yat-sen University, after ethics approval by the ethics board of the Sixth Affiliated Hospital (no. 2022ZSLYEC-236). In addition, we analyzed German patients with rectal cancer from University Hospital RWTH Aachen after ethics approval by the ethics board of the RWTH University Aachen (no. EK028/19). Informed consent was waived for this retrospective study of anonymized data.

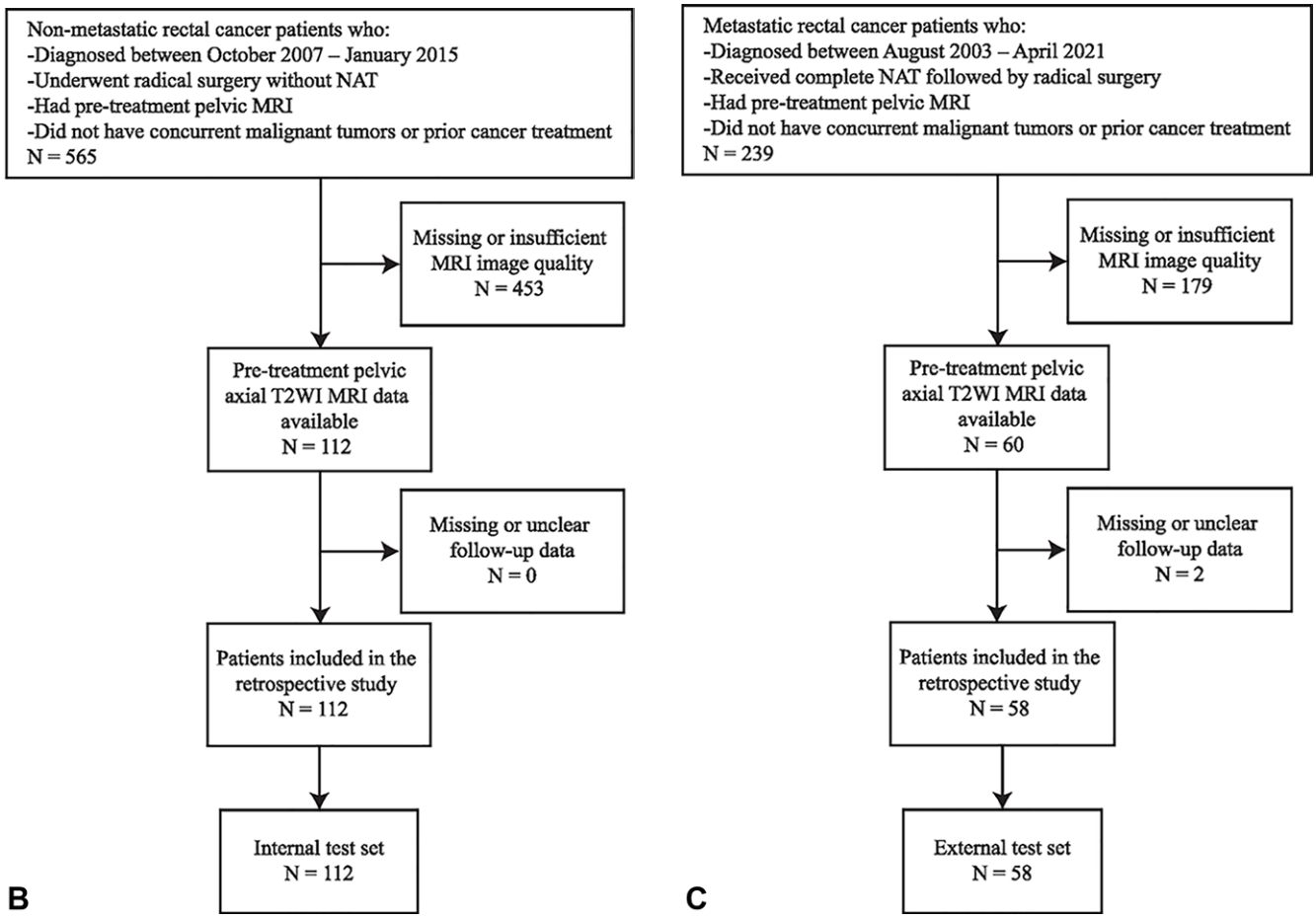

Figure 1 *(continued)*: **(B)** nonmetastatic rectal cancer, which composed the internal test set; and **(C)** metastatic rectal cancer, which composed the external test set. NAT = neoadjuvant therapy, T2WI = T2-weighted imaging.

Patient Cohorts

We collected three independent patient cohorts. First, a primary cohort from China, which was split into a training and a validation set, was used for model selection and hyperparameter optimization. This cohort comprised 725 Chinese patients from the Sixth Affiliated Hospital, Sun Yat-sen University, China, with American Joint Committee on Cancer stage II or III locally advanced rectal cancer diagnosed between June 2010 and April 2019 (Fig 1A). The inclusion criteria were *(a)* histopathologically confirmed rectal adenocarcinoma; *(b)* pretreatment MRI scans available within 1 week before the initiation of neoadjuvant therapy; *(c)* a tumor with a lower edge less than 15 cm above the anal verge as assessed at MRI; *(d)* either pretreatment clinical stage T3 or higher, positive nodal staging, or circumferential resection margin involvement as assessed at MRI; and *(e)* total mesorectal excision performed after neoadjuvant therapy. The exclusion criteria were *(a)* concurrent malignant neoplasms or previous anticancer treatment; *(b)* an incomplete course of neoadjuvant therapy or no radical surgery; *(c)* missing MRI data or insufficient image quality; and *(d)* missing or incomplete clinical follow-up data. Of the 725 patients in the primary cohort, 370 have been previously reported (17). This prior article dealt with the prediction of treatment response after neoadjuvant treatment, whereas in this study, we report on the long-term outcome after treatment.

Second, we collected a nonoverlapping set of 112 patients with rectal cancer diagnosed at the same Chinese center between October 2007 and January 2015 as a completely separate internal test cohort (Fig 1B). The inclusion and exclusion criteria were the same as described earlier, but none of these patients had received neoadjuvant therapy.

Third, we collected data for an external test cohort, which was also only used for the evaluation of the final model (Fig 1C). The data for this cohort were obtained from the University Medical Center of Aachen, Germany, and comprised 58 patients with rectal cancer with the stage IV tumors between August 2003 and April 2021. The inclusion criteria *(a)*, *(b)*, and *(c)* were identical to those of the primary cohort. The inclusion criteria *(d)* and *(e)* were not used in the external test cohort because these were patients with metastases (stage IV tumors). The exclusion criteria were identical to those of the primary cohort.

Image Acquisition and Evaluation

Non–fat-saturated T2-weighted axial MRI scans were obtained (Table S1, Fig 2). In the primary cohort, regions of

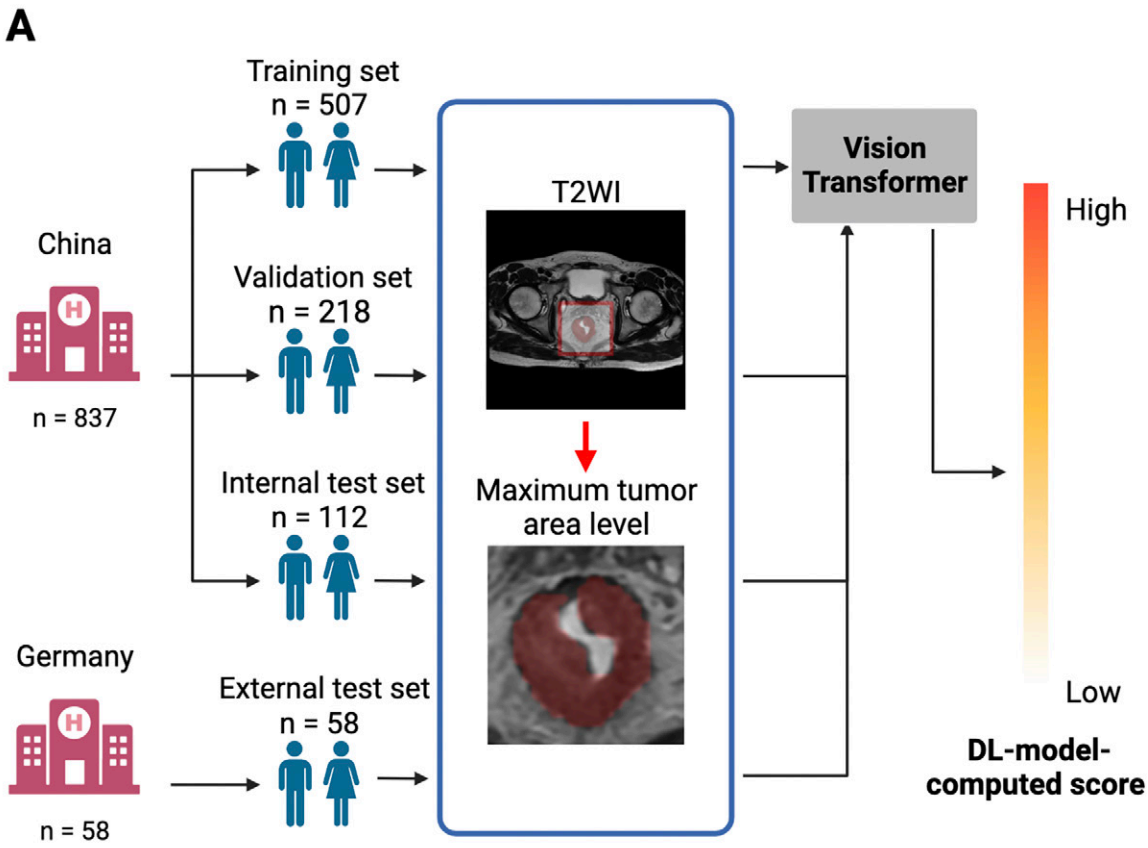

Figure 2: Overview of the workflow of this study. **(A)** Diagram shows the workflow for the development and validation of the deep learning (DL) model. The model was developed using data from 507 patients from the primary cohort as the training set. The model's performance was then validated using the validation set (*n* = 218) and internal test set (*n* = 112) from China and the external test set (*n* = 58) from Germany. All T2-weighted imaging (T2WI) sections were manually expert-delineated, and the section with the maximum tumor area was segmented and then selected as the input to the model. A pretrained Vision Transformer (ViT)–B/16 was used for transfer learning, and the network generated patient-level risk scores *(Fig 2 continues)*.

interest of tumors were manually delineated in a blinded manner in all the T2-weighted sections by two radiologists with 15 years of experience in consensus using ITK-SNAP software (*www.itksnap.org*). In the internal test set and the external test cohort, regions of interest were manually delineated following the same protocol by two surgeons (X.J. and H.Z., with 5 and 3 years of experience, respectively) and were validated by a board-certified radiologist (D.T.) with 10 years of experience in assessing colorectal cancer imaging. All the regions of interest were checked by at least one additional expert with more than 8 years of experience in rectal cancer image processing. The section with the largest tumor region was selected as the input to the model.

DL Procedures and Best Model Selection

We randomly divided eligible patients from the primary cohort into a training set (*n* = 507) and a validation set (*n* = 218) at a 7:3 ratio. We trained a DL model to predict patient risk scores directly from segmented tumor volumes from pretreatment T2-weighted images. The images depict a region of interest centered around a tumor, as well as a 10-pixel area at the outer edge of the tumor. Specifically, we used ViT, which was pretrained on ImageNet-21K (13). We compared this with a more traditional approach, a convolutional neural network (CNN). Specifically, for the ViT, we used the ViT-B/16 architecture. For the CNN, we used the ResNet50 (19), InceptionV3 (20), VGG16 (21), DenseNet121 (22), and MobileNet (23) architecture (Appendix S1). To calculate the difference in survival risk between patients during the optimization process, we used a custom loss function, relying on Cox partial likelihood (24,25). Briefly, two losses for disease-free survival (DFS) and overall survival (OS) were generated, summed at a 1:1 ratio, and used for the gradient calculation to iteratively update the model's parameters in the direction that reduced the loss. The highest Harrell C-index for the validation set was used to select the best checkpoint for each model, and the model with the highest overall C-index was used in subsequent analyses. To visualize which parts of the image were important for prediction, we applied gradient-weighted class activation mapping (26) and an attention map for heatmap generation.

Statistical Analysis

The model's prognostic performance was assessed using the Harrell C-index, time-dependent receiver operating

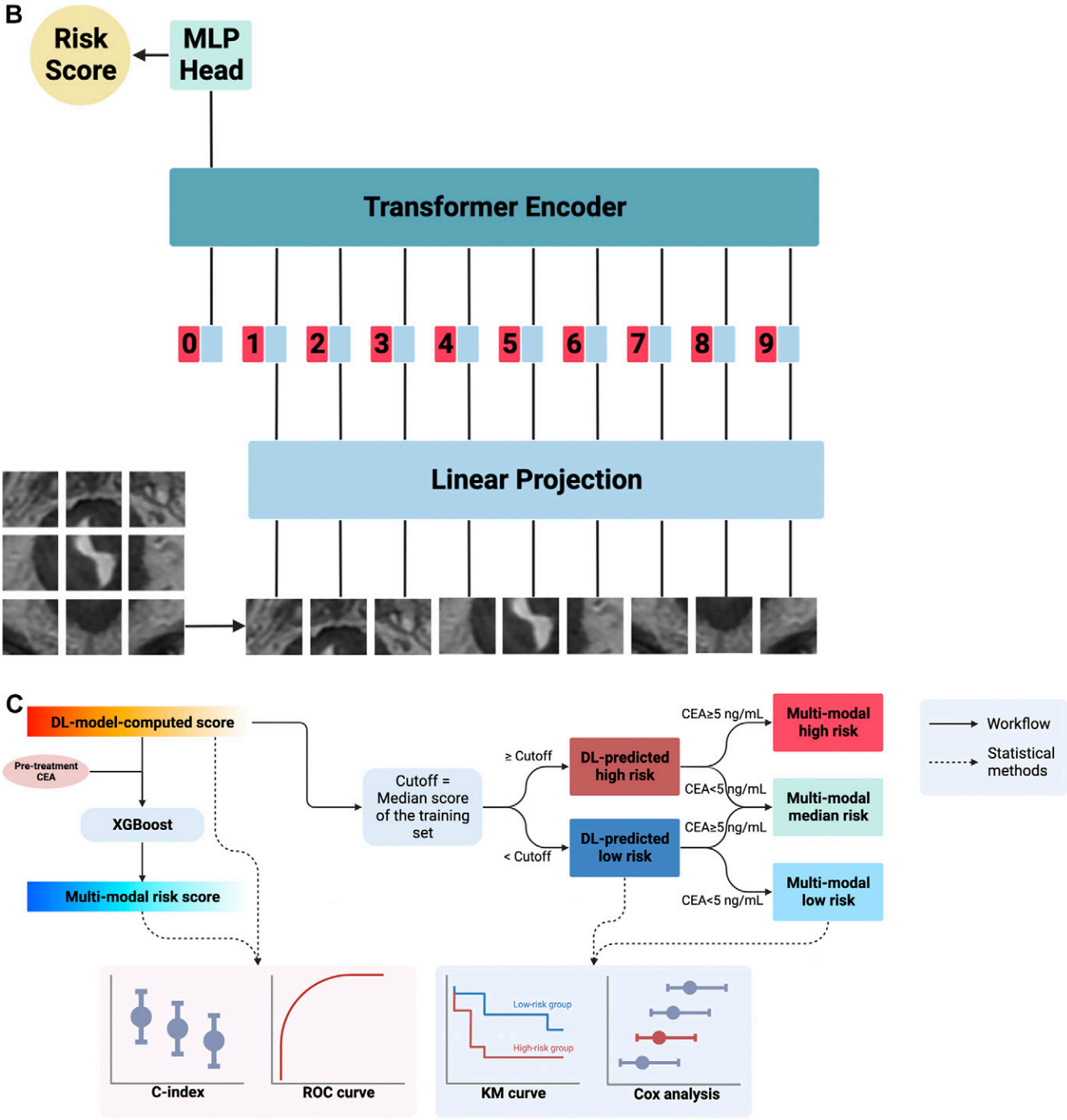

Figure 2 *(continued)*: **(B)** Diagram shows the model structure. MLP = multilayer perceptron. **(C)** Flowchart shows the overall grouping and statistical analysis process, including the assessment of risk scores according to Kaplan-Meier (KM) survival analysis, time-dependent receiver operating characteristic (ROC) analysis, and Cox regression analysis. CEA = carcinoembryonic antigen. Icons reproduced from *https://biorender.com*.

characteristic analyses, and Kaplan-Meier survival analysis. The median risk score in the training cohort was used to stratify patients in each data set; those with a risk score above or equal to −1.307 were placed in the high-risk group, and those with a risk score below this value were placed in the low-risk group. Kaplan-Meier survival analysis was performed to assess the difference in prognosis between these two groups. Post hoc analyses were conducted to determine the relationship between two groups and subgroups split by sex, age, tumor location, T stage, N stage, pretreatment carcinoembryonic antigen (CEA) level (in nanograms per milliliter), and radiation therapy.

Statistical analyses were performed using R, version 3.4.0 (R Foundation), by X.J. and H.Z. Patient characteristics were compared using the *t* test for continuous variables and χ^2 or Fisher exact test for categorical variables. OS was defined as the time between surgery and death from any cause. DFS was defined as the interval from surgery to recurrence

Table 1: Patient Characteristics in the Training, Validation, Internal Test, and External Test Sets

Variable	Training Set (n = 507)	Validation Set (n = 218)	P Value	Internal Test Set (n = 112)	External Test Set (n = 58)
Sex			.33		
F	152 (30.0)	74 (33.9)		36 (32.1)	20 (34.5)
M	355 (70.0)	144 (66.1)		76 (67.9)	38 (65.5)
Age			.51		
≤50 years	189 (37.3)	72 (33.0)		27 (24.1)	16 (27.6)
51–65 years	225 (44.4)	101 (46.3)		50 (44.6)	25 (43.1)
>65 years	93 (18.3)	45 (20.6)		35 (31.3)	17 (29.3)
Pretreatment CEA level (ng/mL)*	3.9 (1.9–9.2)	3.8 (2.1–9.5)	.22	2.6 (1.3–4.4)	24.9 (9.7–76.5)
No. with missing data	0	0		6 (5.4)	26 (44.8)
Radiation therapy			.82		
No	246 (48.5)	103 (47.2)		NA	17 (29.3)
Yes	261 (51.5)	115 (52.8)		NA	41 (70.7)
pT stage			.74		
T0	114 (22.5)	50 (22.9)		0	NA
T1	31 (6.1)	19 (8.7)		21 (18.8)	NA
T2	110 (21.7)	42 (19.3)		31 (27.7)	4 (6.9)
T3	250 (49.3)	106 (48.6)		58 (51.8)	39 (67.2)
T4	2 (0.4)	1 (0.5)		2 (1.8)	5 (8.6)
Missing	0	0		0	10 (17.2)
pN stage			.07		
N0	377 (74.4)	170 (78.0)		84 (75.0)	13 (22.4)
N1	94 (18.5)	42 (19.3)		21 (18.8)	16 (27.6)
N2	36 (7.1)	6 (2.8)		7 (6.3)	18 (31.0)
Missing	0	0		0	11 (19.0)
OS (y)*	2.6 (1.5–4.0)	2.8 (1.7–4.4)		4.6 (3.8–5.8)	1.7 (1.1–3.7)
DFS (y)*	2.4 (1.2–3.9)	2.3 (1.4–4.3)		4.6 (3.3–5.8)	0.6 (0.3–1.0)

Note.—Unless stated otherwise, data are numbers of patients, with percentages in parentheses. Pretreatment T and N stages were determined at MRI, while pT and pN refer to pathologic stages. P values were calculated using the t test for continuous variables and the χ^2 test for categorical variables. CEA = carcinoembryonic antigen, DFS = disease-free survival, NA = not applicable, OS = overall survival.

* Data are medians, with IQRs in parentheses.

or metastasis or death from any cause. All statistical tests were twosided, and $P < .05$ was considered indicative of statistically signif icant difference.

Code Availability

All source code and trained models are publicly available at *https://github.com/KatherLab/dl-mri*.

Results

Patient Demographic Characteristics

For the primary cohort, among 2133 patients with resected locally advanced rectal cancers, we excluded 1303 patients due to the absence of a pretreatment pelvic axial T2-weighted MRI scan or insufficient image quality, as well as 105 patients whose survival data were missing. For the internal test set, 565 patients with resected nonmetastatic rectal cancer were included, with 453 patients then excluded due to the absence of a pretreatment pelvic axial T2-weighted MRI scan or insufficient image quality. For the external test cohort, 239 patients with resected metastatic rectal cancer were included, with 179 patients then excluded due to the absence of a pretreatment pelvic axial T2-weighted MRI scan or insufficient image quality and two patients excluded due to missing survival data. The characteristics of patients in the training set (n = 507), validation set (n = 218), internal test set (n = 112), and external test set (n = 58) are summarized in Table 1. The median age was 56 years (IQR, 46–64 years) in the training set, 55 years (IQR, 47–63 years) in the validation set, 60 years (IQR, 52–70 years) in the internal set, and 57 years (IQR, 50–67 years) in the external test set. No statistically significant difference in baseline information between the training and internal test set was observed (Table 1).

DL Predicts Outcomes in the Validation Set

We trained a ViT model to predict survival from segmented tumor volumes on pretreatment T2-weighted images. The ViT model had C-indexes of 0.82 for OS and 0.69 for DFS, which were higher than the C-indexes calculated for any of the CNNs tested (Fig S1, Table S2). Using the validation set, the risk score output from the ViT model was able

Table 2: Characteristics of Patients Stratified by High and Low Risk in the Training, Validation, Internal Test, and External Test Sets

	Training Set			Validation Set			Internal Test Set			External Test Set		
Variable	High Risk (*n* = 253)	Low Risk (*n* = 254)	*P* Value	High Risk (*n* = 108)	Low Risk (*n* = 110)	*P* Value	High Risk (*n* = 69)	Low Risk (*n* = 43)	*P* Value	High Risk (*n* = 31)	Low Risk (*n* = 27)	*P* Value
Sex			.22			.23			.89			.65
F	69 (27.3)	83 (32.7)		32 (29.6)	42 (38.2)		23 (33.3)	13 (30.2)		12 (38.7)	8 (29.6)	
M	184 (72.7)	171 (67.3)		76 (70.4)	68 (61.8)		46 (66.7)	30 (69.8)		19 (61.3)	19 (70.4)	
Age			.54			.89			.83			.06
≤50 years	94 (37.2)	95 (37.4)		34 (31.5)	38 (34.5)		18 (26.1)	9 (20.9)		6 (19.4)	10 (37.0)	
51–65 years	108 (42.7)	117 (46.1)		51 (47.2)	50 (45.5)		30 (43.5)	20 (46.5)		12 (38.7)	13 (48.1)	
>65 years	51 (20.2)	42 (16.5)		23 (21.3)	22 (20.0)		21 (30.4)	14 (32.6)		13 (41.9)	4 (14.8)	
pT stage			.16			.04			.86			>.99
T0, T1, or T2	119 (47.0)	136 (53.5)		47 (43.5)	64 (58.2)		33 (47.8)	19 (44.2)		2 (6.5)	2 (7.4)	
T3 or T4	134 (53.0)	118 (46.5)		61 (56.5)	46 (41.8)		36 (52.2)	24 (55.8)		22 (71.0)	22 (81.5)	
Missing	0	0		0	0		0	0		7 (22.6)	3 (11.1)	
pN stage			<.001			.01			.11			.57
N0	170 (67.2)	207 (81.5)		76 (70.4)	94 (85.5)		55 (79.7)	29 (67.4)		5 (16.1)	8 (29.6)	
N+	83 (32.8)	47 (18.5)		32 (29.7)	16 (14.5)		14 (20.3)	14 (32.6)		18 (58.0)	16 (59.3)	
Missing	0	0		0	0		0	0		8 (25.8)	3 (11.1)	

Note.—Unless stated otherwise, data are numbers of patients, with percentages in parentheses. Patients were stratified by a high- or low-risk deep learning (DL) model–computed score. A cutoff value of −1.307 for the DL model–computed score was used, which was determined in the training set only. *P* values were calculated using the *t* test for continuous variables and the χ^2 test for categorical variables. Pretreatment T and N stages were determined at MRI, while pT and pN refer to pathologic stages.

to accurately predict 1-, 3-, and 5-year OS, with an area under the receiver operating characteristic curve (AUC) of 0.85 (95% CI: 0.77, 0.94), 0.90 (95% CI: 0.82, 0.97), and 0.79 (95% CI: 0.62, 0.97), respectively. The DL model–computed risk score was also able to accurately predict DFS at 1 year (AUC, 0.74 [95% CI: 0.61, 0.86]), 3 years (AUC, 0.70 [95% CI: 0.58, 0.82]), and 5 years (AUC, 0.70 [95% CI: 0.57, 0.83]) in the validation set (Fig S2). Patients were also classified into high-risk and low-risk groups based on DL model–computed risk score, and clinical characteristics for these two groups are shown in Table 2. In both the training and validation sets, the number of lymph node metastases was greater in the high-risk group than in the low-risk group. Patients in the high-risk group had decreased OS (hazard ratio [HR], 7.1 [95% CI: 2.4, 21.3]; $P < .001$) and DFS (HR, 2.7 [95% CI: 1.3, 5.4]; $P = .005$) compared with patients in the low-risk group in the validation set (Fig 3A).

The DL model–computed risk score distribution of the validation set as well as example gradient-weighted class activation mapping and attention maps of the original images are shown in Figure 3B.

Model Predictions Generalize to Separate Cohorts and Different Clinical Subgroups

We validated the model in two separate test cohorts. The data from the first test cohort (ie, internal test cohort) were obtained from the same medical center as those in the training cohort and included pretreatment T2-weighted MRI scans from 112 patients with nonmetastatic rectal cancer. In the internal test cohort, the C-index for OS and DFS was 0.65 and 0.60, respectively. The other test cohort (ie, external test cohort) included 58 patients with metastatic rectal cancer from Germany. In this external test cohort, the C-index was 0.62 and 0.54 for OS and DFS, respectively. No difference was observed for DFS between the DL model–computed high- and low-risk groups in the external test cohort ($P = .25$), but the DL model–predicted high-risk group showed decreased OS, with a HR of 2.3 (95% CI: 1.0, 5.4; $P = .047$) (Figs 4, S3). Subgroup analyses performed on the primary cohort, as well as the internal test set, showed that DL model–predicted risk was a strong predictor in almost all clinical subgroups (Figs S4, S5; Table S3). Even in subgroups that are generally considered to have a better prognosis, such as those with tumor regression grading 0 or 1, the DL model–computed risk score was able to distinguish between high- and low-risk groups in the validation set (log-rank $P = .04$) (Fig S6). In the primary cohort, radiation therapy did not improve OS ($P = .95$) (Fig S7A) but did improve DFS and reduce recurrence rates (13.7% vs 25.2%; $P = .007$) (Fig S7D). Compared with the high-risk group, the low-risk group showed improved DFS following radiation therapy (HR, 0.6 [95% CI: 0.3, 1.0]; $P = .04$) (Fig S7E, S7F). No difference in survival was observed between patients who did and did not receive radiation therapy in the DL model–computed high-risk group.

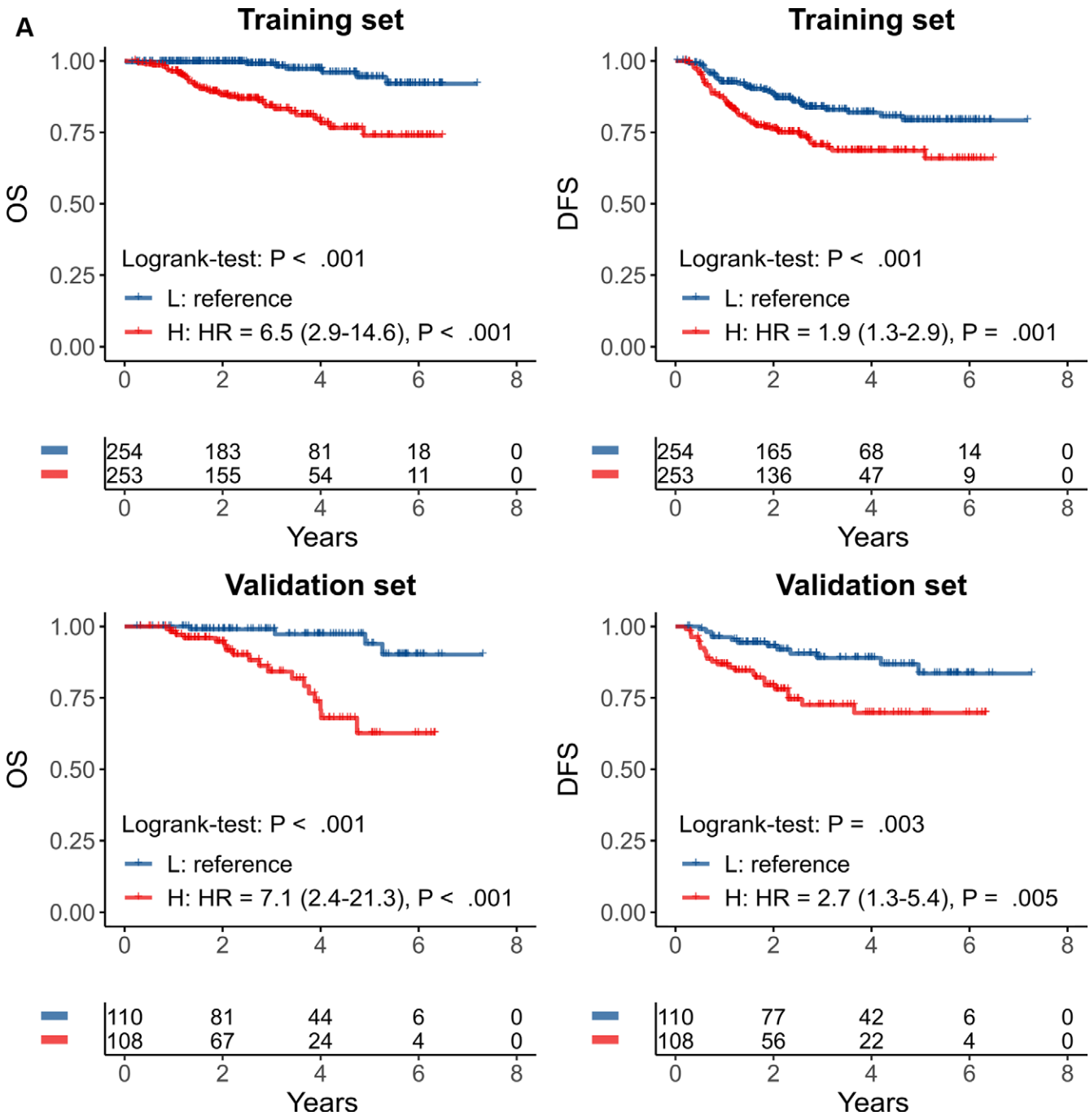

Figure 3: Kaplan-Meier curves show deep learning (DL) model–computed risk score based on the training set and validation set accompanied by distribution and visualization of the DL model–computed risk score in the validation set. **(A)** Kaplan-Meier curves for overall survival (OS) and disease-free survival (DFS) in the training set (top) (n = 507) and validation set (bottom) (n = 218) with patients stratified by a high-risk (H) (red line) or low-risk (L) (blue line) DL model–computed risk score. A cutoff value of −1.307 for the DL signature was used, which was determined in the training set only. The log-rank test was used to calculate P values. The P values given after hazard ratios (HRs) represent the significance test of the Cox regression, for which the likelihood ratio test was used. Data in parentheses are 95% CIs *(Fig 3 continues)*.

Model Predictions Are Statistically Independent of Potential Confounders

As shown in Figures 5 and S8, for OS, the multivariable Cox regression–adjusted HRs of the DL model–predicted risk were 6.0 (95% CI: 2.6, 13.8; P < .001), 4.6 (95% CI: 1.4, 15.2; P = .01), and 3.2 (95% CI: 1.0, 9.8; P = .04) in the training, validation, and internal test sets, respectively, and were higher than for the pathologic T or N stage. A comparison of the C-index of these variables showed similar results (Table S4). For DFS, the HRs in the training, validation, and internal test sets were 1.6 (95% CI: 1.0, 2.5; P = .03), 1.6 (95% CI: 0.8, 3.4; P = .21), and 2.8 (95% CI: 1.1, 6.8; P = .03), respectively.

Multimodal Input Models Improve Performance

The integrated XGBoost model achieved a higher C-index of 0.86 (95% CI: 0.79, 0.93) for OS compared with DL

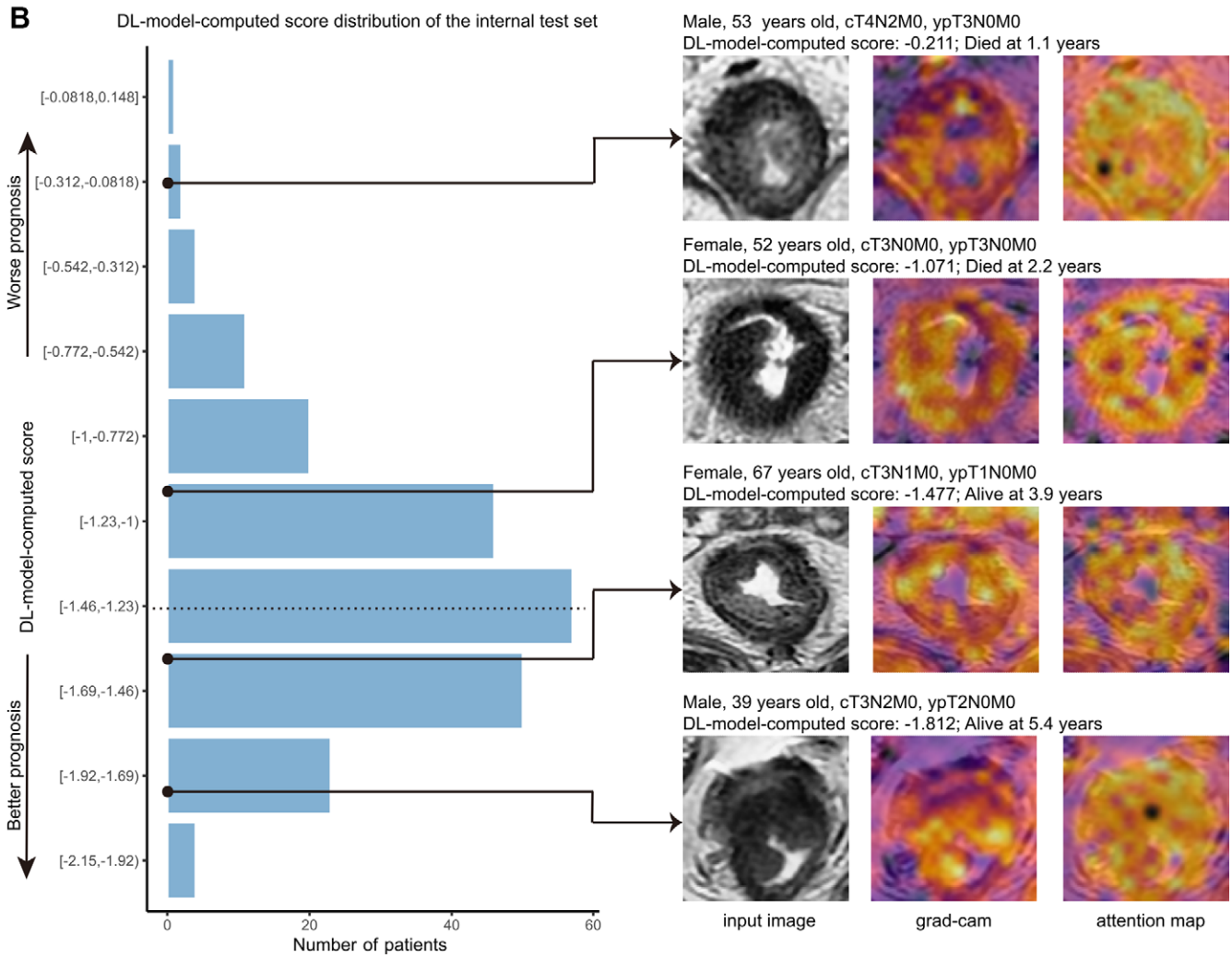

Figure 3 *(continued)*: **(B)** A histogram of the distribution of DL model–computed risk score is shown on the left. The dashed line represents the cutoff value. The right side of the figure shows examples of segmented MRI scans with the expert-delineated region of interest that were used as input for Vision Transformer (ViT). The gradient-weighted class activation map (grad-cam) heat map and attention map generated by ViT, overlaid on the input image, are shown on the right side of the input image. The ViT model demonstrated a high level of attention toward the tumor region in the images.

model–computed risk score alone (C-index, 0.82 [95% CI: 0.74, 0.90]) and pretreatment CEA levels alone (C-index, 0.61 [95% CI: 0.48, 0.74]) in the validation set. While the multimodal model showed some improvement in prognostication in both the internal and external test sets, it only significantly enhanced prognostic accuracy in the external test set (P = .045) (Fig 6A). Despite this, both the DL model–computed risk score and the multimodal model performed significantly better than pretreatment CEA alone in terms of prognostication (Table S5). Further, to facilitate clinical application, we combined CEA status (<5 ng/mL or ≥5 ng/mL) with DL model–predicted risk groups and divided the patients into three new multimodal risk groups, as illustrated in Figure 2C. The multimodal low-risk group had the most favorable prognosis compared with the multimodal high-risk group. In the internal test set, the HR was 4.5 (95% CI: 1.0, 20.1; P = .049) for OS and 3.4 (95% CI: 1.0, 11.7; P = .05) for DFS (Fig 6B).

Discussion

We hypothesized that prognostic information can be extracted from MRI at the time of rectal cancer diagnosis with use of deep learning (DL) to help optimize treatment and monitoring. Herein, we developed a Vision Transformer (ViT)–based DL model for predicting the survival of patients with rectal cancer from pelvic MRI data. Our ViT-based model achieved considerable predictive performance (hazard ratio [HR] of the training set, 6.5 [95% CI: 2.9, 14.6] [P < .001] and HR of validation set, 7.1 [95% CI: 2.4, 21.3] [P < .001]) in the primary cohort, which included patients with TNM stage II or III rectal cancer. We validated the model in an internal test set with patients with nonmetastatic rectal cancer and an external test cohort that included patients with TNM stage IV rectal cancer, and our model was still able to predict prognosis well (HR of internal test set, 3.0 [95% CI: 1.0, 9.0] [P = .047] and HR of external test cohort, 2.3 [95% CI: 1.0, 5.4] [P = .047]). In addition, when adjusted for clinicopathologic factors such

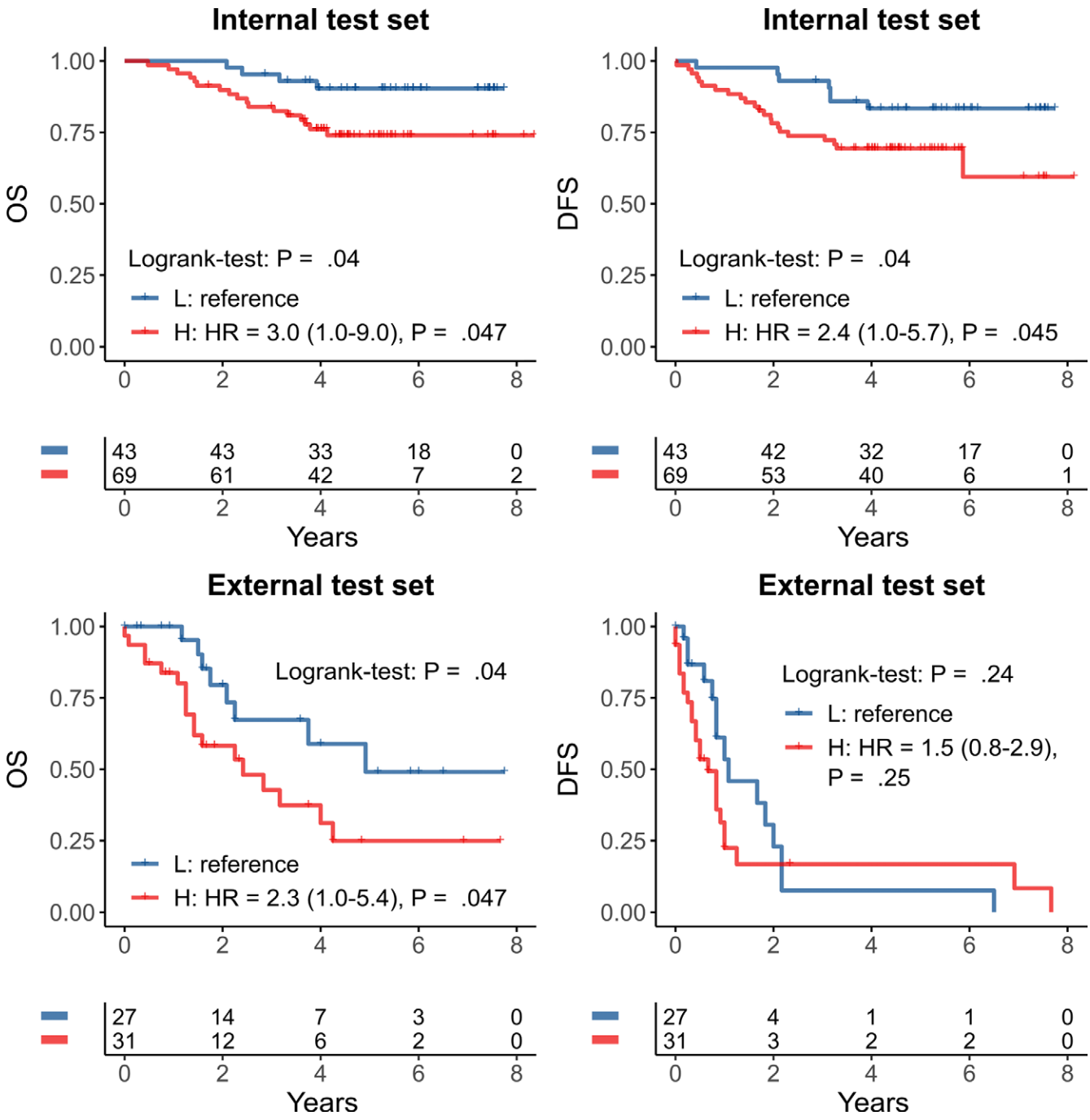

Figure 4: Kaplan-Meier curves show deep learning (DL) model–computed risk score based on the internal and external test sets. Kaplan-Meier curves for overall survival (OS) and disease-free survival (DFS) in the internal set (top) (n = 112) and external test set (bottom) (n = 58) with patients stratified by a high-risk (H) (red line) and low-risk (L) (blue line) DL model–computed risk score. A cutoff value of −1.307 for the DL model–computed risk score was used, which was determined in the training set only. The log-rank test was used to calculate P values. The P values given after hazard ratios (HRs) represent the significance test of the Cox regression, for which the likelihood ratio test was used. Data in parentheses are 95% CIs.

as TNM stage and carcinoembryonic antigen levels, the DL model–predicted risk score remained an independent predictor of patient outcomes.

Previous survival prediction studies were based on classic radiomics and achieved an AUC or C-index of approximately 0.8 at internal validation in single-center settings (27–29), whereas our transformer-based model not only demonstrated similar performance, but has also been validated in a larger sample size and across different countries. Additionally, radiomics usually requires precise manual tumor annotation and empirical feature extraction, which limits the reproducibility and scalability of the study. In contrast, in this study, we used single-level tumor regions as input, which could greatly reduce the workload of subsequent studies. In addition, survival prediction research, unlike image classification, requires an accurate ordering

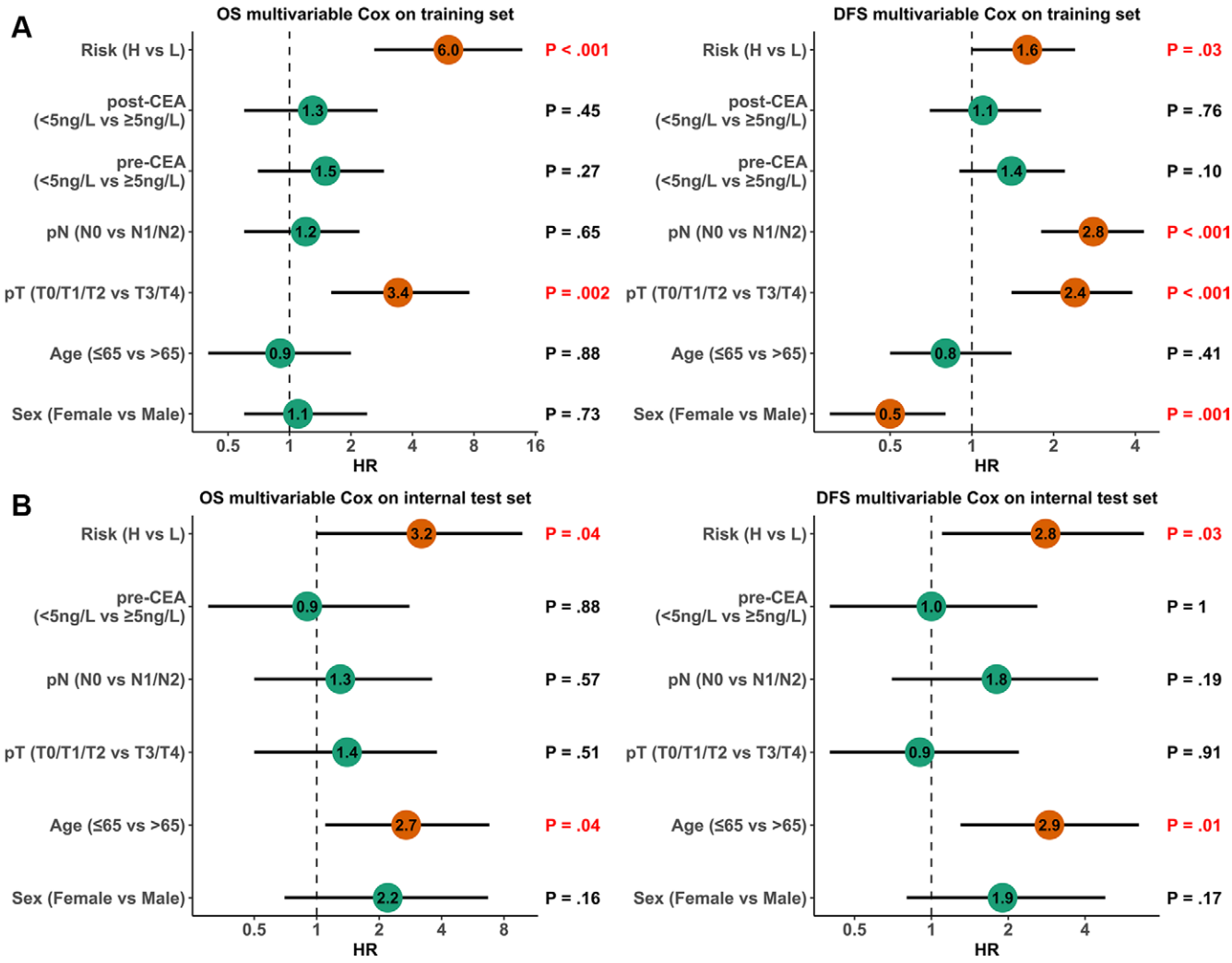

Figure 5: Forest plots for the multivariable Cox regression analysis. **(A)** Multivariable Cox regression analysis between deep learning (DL) model–predicted risk and clinical features for overall survival (OS) and disease-free survival (DFS) in the training set. **(B)** Multivariable Cox regression analysis between DL model–predicted risk and clinical features for OS and DFS in the internal test set. The horizontal axis is scaled logarithmically ($\log_2$), and the vertical dotted line represents hazard ratios (HRs) equal to 1. Numbers inside the circles are HRs. Green circles represent $P > .05$, orange circles represent $P \leq .05$, and horizontal lines on either side of the circles represent 95% CIs. H = high risk, L = low risk, pN = pathologic N stage, post-CEA = carcinoembryonic antigen level after neoadjuvant therapy, pre-CEA = carcinoembryonic antigen level before neoadjuvant therapy or surgery (for the internal test set), pT = pathologic T stage.

of survival time. However, follow-up data often have censored data, which makes it difficult to use DL for survival prediction. While previous studies have typically adopted binary classification to overcome this issue (30,31), in this study, we used Cox partial likelihood as the loss function to calculate the prognostic risk of the patient in a continuous way. In so doing, we reduced the subjective effects of setting an arbitrary cutoff value to binarize survival time. Furthermore, we incorporated loss functions for both OS and DFS into the training, enabling the DL model to learn the features of recurrence, metastasis, and death, rather than just a single event.

We used ViT-B/16 as the network backbone in this study, which has emerged as an alternative to CNNs in nonmedical computer vision tasks (13). Although ViT has achieved better performance than traditional CNNs in these tasks, to our knowledge, it has not yet been widely used in medical image classification, let alone survival prediction (32,33). Previous studies have shown that transformers such as ViT-B/16 are more robust to high-frequency noise in image data (34,35). In this study, we empirically demonstrated that the pretrained ViT can achieve better performance than traditional CNNs in terms of survival prediction (OS C-index, 0.82 vs 0.67 [ResNet]; $P = .02$).

The watch-and-wait approach is currently a safe strategy for patients with rectal cancer (36). However, identifying patients with high risk after they have achieved clinical complete response remains challenging. In subgroups that usually have better prognostic outcomes—such as those with tumor regression grading 0 and 1 and N0 (37,38)—the DL model–predicted risk can distinguish between high- and low-risk patients. This suggests that our DL model–computed risk score could help identify high-risk patients who may benefit from more aggressive treatment even if they

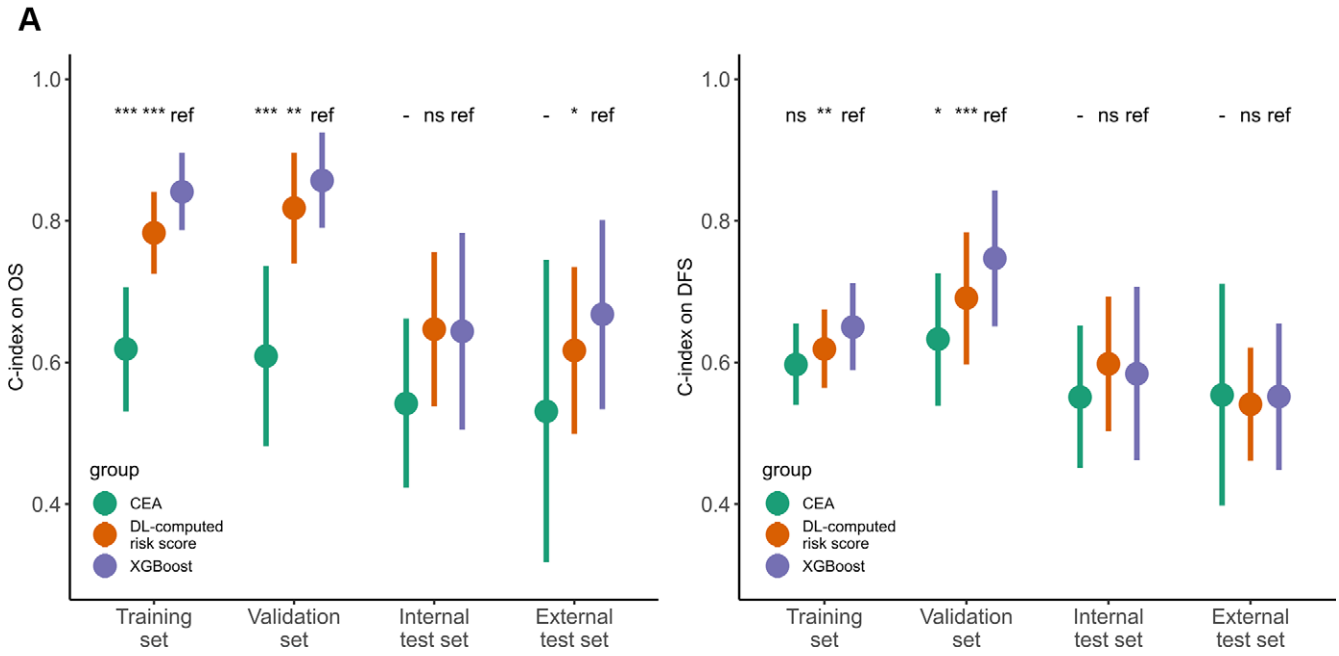

Figure 6: Integration of deep learning (DL) model–computed risk score and a blood-based tumor biologic marker. **(A)** Comparison of Harrell C-indexes for pretreatment carcinoembryonic antigen (CEA) level in nanograms per milliliter (green), DL model–computed risk score (orange), and XGBoost model (purple) (combination of pretreatment CEA level and DL model–computed risk score into one model) in the training set ($n = 507$), validation set ($n = 218$), internal test set ($n = 112$), and external test set ($n = 58$) for overall survival (OS) and disease-free survival (DFS). Circles represent the C-index, and lines on either side of the circle represent the 95% CI. CEA and DL model–computed risk scores are compared with XGBoost as the reference. ns = not significant, ref = reference group. Not significant = $P \geq 0.05$; * = $.01 \leq P < .05$; ** = $.001 \leq P < .01$; *** = $P < .001$ *(Fig 6 continues).*

have a good response after neoadjuvant therapy. This could be a useful addition to the decision-making process for the watch-and-wait approach. We also found that the DL model–predicted high-risk group did not show improved survival following radiation therapy (log-rank $P = .18$). This suggests that our DL model–computed risk score could be used to identify patients who would not benefit from additional radiation therapy treatment. However, considering the effects of sample size and confounding factors, this conclusion still needs further validation.

We also analyzed a multimodal approach that included pretreatment CEA levels, a blood biomarker whose high concentration has been shown to be associated with poor prognosis (39). Our results show that multimodal models have better efficacy than DL model–computed risk score alone (C-index, 0.75 vs 0.69 [$P < .001$] in the validation set), and future studies may consider adding more prognostic-related factors to build a more robust model.

Our study had several limitations. First, although we validated the model in a cross-national cohort, the clinical stage as well as treatment regimens were not consistent between the external test set and the training cohort, and the cutoff used for high- and low-risk stratification was based on normalization within the sample. Moreover, the external test set consisted of patients with advanced disease, which may have limited the potential for management changes. In addition, all data were retrospectively collected, and the sample sizes of the internal and external test sets were small, which may introduce inherent bias and hidden confounding factors. Thus, a larger prospective cohort is needed for validation. Nevertheless, we show that our prognostic model upholds its performance in external validation even though the clinical characteristics of the individual cohorts are different. This represents a robust generalization and is a promising indication that the model genuinely detects prognostic morphologic factors in rectal cancer.

In conclusion, we developed and validated a preoperative MRI–based deep learning model to predict survival of patients with rectal cancer without any manual feature extraction or selection. The output of the model is a risk score with independent prognostic utility and could be used as a preoperative risk stratification tool for patients with rectal cancer. Our study shows that transformer neural networks should be considered a promising alternative to convolutional neural networks for radiology image analysis. Our open-source model could contribute to advancing precision medicine and warrants further validation in future clinical trials.

Author contributions:: Guarantors of integrity of entire study, **X.J., H.Z., J.N.K., J.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting, **X.J.**; 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.J., O.L.S.**; clinical studies, **X.J., H.Z., X.W., X.M., J.K.**; experimental studies, **X.J., O.L.S., J.N.K.**; statistical analysis, **X.J., H.Z., J.N.K., J.K.**; and manuscript editing, **X.J., H.Z., O.L.S., S.N., C.K., I.A., S.A.L., X.W., D.T., J.N.K., J.K.**

Disclosures of conflicts of interest: X.J. No relevant relationships. **H.Z.** No relevant relationships. **O.L.S.** No relevant relationships. **S.N.** No relevant relationships. **C.K.** Participation on a data safety monitoring board or advisory board for Bayer and Guerbet. **I.A.** No relevant relationships. **S.A.L.** No relevant relationships. **X.W.** No relevant relationships. **X.M.** No relevant relationships. **D.T.** No relevant relationships. **J.N.K.** Consulting fees from Owkin, France

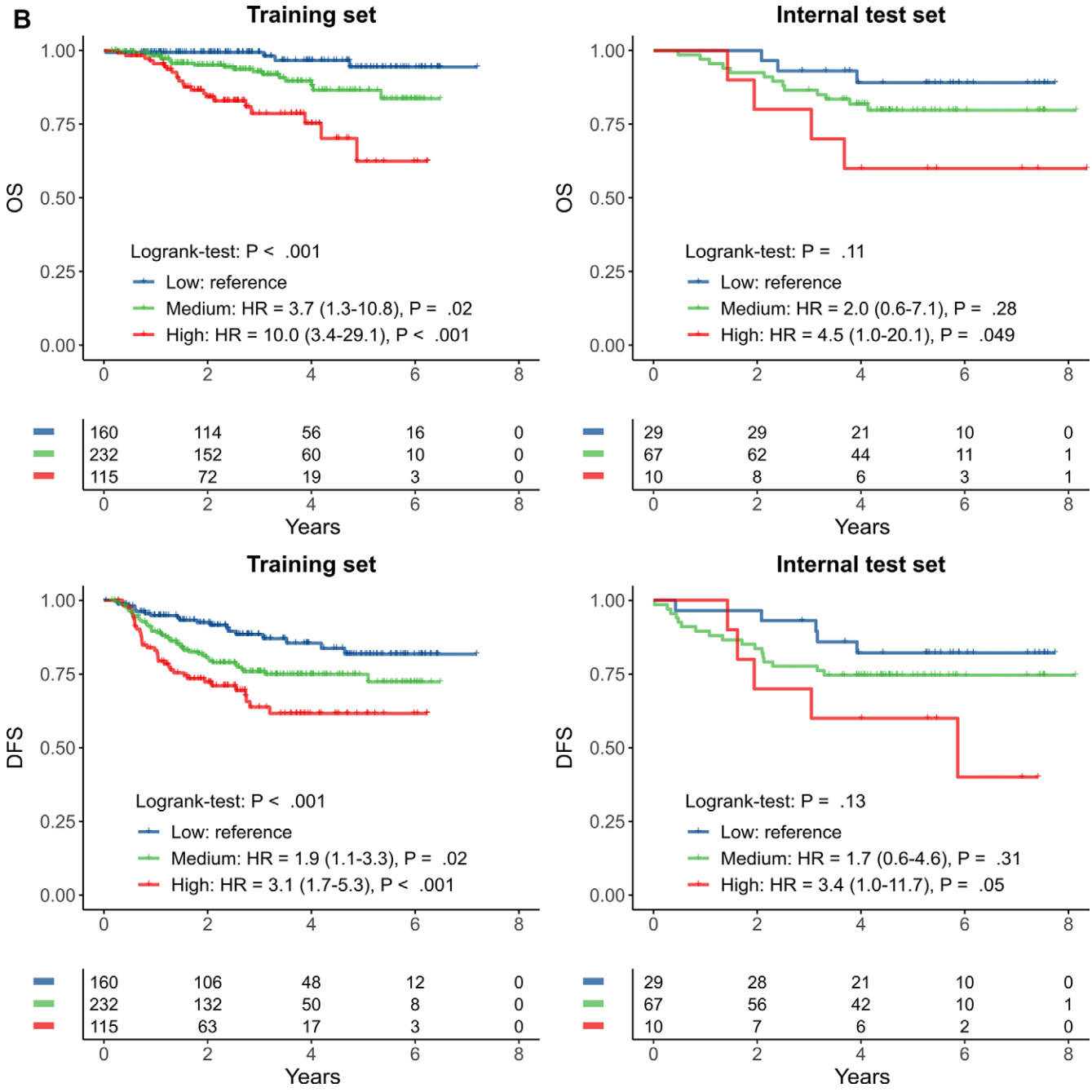

Figure 6 *(continued)*: **(B)** Kaplan-Meier curves in patients stratified into multimodal low-risk (green line), medium-risk (blue line), and high-risk (red line) groups based on a combination of DL model–computed risk score and pretreatment CEA level. Kaplan-Meier curves for OS are shown for the training set in top left and the internal test set in top right. The bottom row shows the Kaplan-Meier curves for DFS for the training and internal test sets. The *P* values given after hazard ratios (HRs) represent the significance test of the Cox regression, for which the likelihood ratio test was used. Data in parentheses are 95% CIs.

Panakeia, UK DoMore Diagnostics, and Oslo; payment for lectures form Fresenius, Eisai, and MSD; support for travel from the European Society of Medical Oncology. **J.K.** No relevant relationships.

References

1. Bosset JF, Collette L, Calais G, et al. Chemotherapy with preoperative radiotherapy in rectal cancer. N Engl J Med 2006;355(11):1114–1123. [Published correction appears in N Engl J Med 2007;357(7):728.]
2. Azria D, Doyen J, Jarlier M, et al. Late toxicities and clinical outcome at 5 years of the ACCORD 12/0405-PRODIGE 02 trial comparing two neoadjuvant chemoradiotherapy regimens for intermediate-risk rectal cancer. Ann Oncol 2017;28(10):2436–2442.
3. Rödel C, Liersch T, Becker H, et al. Preoperative chemoradiotherapy and postoperative chemotherapy with fluorouracil and oxaliplatin versus fluorouracil alone in locally advanced rectal cancer: initial results of the German CAO/ARO/AIO-04 randomised phase 3 trial. Lancet Oncol 2012;13(7):679–687.
4. Peeters KCMJ, Marijnen CAM, Nagtegaal ID, et al. The TME trial after a median follow-up of 6 years: increased local control but no survival benefit in irradiated patients with resectable rectal carcinoma. Ann Surg 2007;246(5):693–701.
5. Sebag-Montefiore D, Stephens RJ, Steele R, et al. Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial. Lancet 2009;373(9666):811–820.

6. Fokas E, Liersch T, Fietkau R, et al. Tumor regression grading after preoperative chemoradiotherapy for locally advanced rectal carcinoma revisited: updated results of the CAO/ARO/AIO-94 trial. J Clin Oncol 2014;32(15):1554–1562.
7. Wibe A, Rendedal PR, Svensson E, et al. Prognostic significance of the circumferential resection margin following total mesorectal excision for rectal cancer. Br J Surg 2002;89(3):327–334.
8. Lord AC, D'Souza N, Shaw A, et al. MRI-diagnosed tumor deposits and EMVI status have superior prognostic accuracy to current clinical TNM staging in rectal cancer. Ann Surg 2022;276(2):334–344.
9. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–446.
10. Cui Y, Yang X, Shi Z, et al. Radiomics analysis of multiparametric MRI for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Eur Radiol 2019;29(3):1211–1220.
11. Liu Z, Meng X, Zhang H, et al. Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. Nat Commun 2020;11(1):4308.
12. Pei Q, Yi X, Chen C, et al. Pre-treatment CT-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. Eur Radiol 2022;32(1):714–724.
13. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv 2010.11929 [preprint] https://arxiv.org/abs/2010.11929. Posted October 22, 2020. Accessed September 2022.
14. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology 2019;290(2):290–297.
15. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. J Pathol 2022;257(4):430–444.
16. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–510.
17. Jin C, Yu H, Ke J, et al. Predicting treatment response from longitudinal images using multi-task deep learning. Nat Commun 2021;12(1):1851.
18. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015;102(3):148–158.
19. 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.
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv 1512.00567 [preprint] https://arxiv.org/abs/1512.00567. Posted December 2, 2015. Accessed January 9, 2023.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556 [preprint] https://arxiv.org/abs/1409.1556. Posted September 4, 2014. Accessed September 2022.
22. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 4700–4708.
23. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv 1704.04861 [preprint] https://arxiv.org/abs/1704.04861. Posted April 17, 2017. Accessed September 2022.
24. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018;18(1):24.
25. Jiang Y, Jin C, Yu H, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. Ann Surg 2021;274(6):e1153–e1161.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv 1610.02391 [preprint] https://arxiv.org/abs/1610.02391. Posted October 7, 2016. Accessed January 9, 2023.
27. Nie K, Hu P, Zheng J, et al. Incremental value of radiomics in 5-year overall survival prediction for stage II-III rectal cancer. Front Oncol 2022;12:779030.
28. Cui Y, Wang G, Ren J, et al. Radiomics features at multiparametric MRI predict disease-free survival in patients with locally advanced rectal cancer. Acad Radiol 2022;29(8):e128–e138.
29. Ghaffari Laleh N, Ligero M, Perez-Lopez R, Kather JN. Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer. Clin Cancer Res 2023;29(2):316–323.
30. Li CH, Cai D, Zhong ME, et al. Multi-size deep learning based preoperative computed tomography signature for prognosis prediction of colorectal cancer. Front Genet 2022;13:880093.
31. Liu X, Zhang D, Liu Z, et al. Deep learning radiomics-based prediction of distant metastasis in patients with locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a multicentre study. EBioMedicine 2021;69:103442.
32. Mondal AK, Bhattacharjee A, Singla P, Prathosh AP. xViTCOS: explainable Vision Transformer based COVID-19 screening using radiography. IEEE J Transl Eng Health Med 2021;10:1100110.
33. Wu Y, Qi S, Sun Y, Xia S, Yao Y, Qian W. A vision transformer for emphysema classification using CT images. Phys Med Biol 2021;66(24):245016.
34. Ghaffari Laleh N, Truhn D, Veldhuizen GP, et al. Adversarial attacks and adversarial robustness in computational pathology. Nat Commun 2022;13(1):5711.
35. Ghiasi A, Kazemi H, Borgnia E, et al. What do vision transformers learn? A visual exploration. arXiv 2212.06727 [preprint] https://arxiv.org/abs/2212.06727. Posted December 13, 2022. Accessed September 2022.
36. Jayaprakasam VS, Alvarez J, Omer DM, Gollub MJ, Smith JJ, Petkovska I. Watch-and-wait approach to rectal cancer: the role of imaging. Radiology 2023;307(1):e221529.
37. Fokas E, Ströbel P, Fietkau R, et al. Tumor regression grading after preoperative chemoradiotherapy as a prognostic factor and individual-level surrogate for disease-free survival in rectal cancer. J Natl Cancer Inst 2017;109(12):.
38. 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.
39. Duffy MJ. Carcinoembryonic antigen as a marker for colorectal cancer: is it clinically useful? Clin Chem 2001;47(4):624–630.