Supplemental Appendix A. Detailed Description of Treatment Regimen and Image Acquisition

Treatment Regimen

All patients were treated with external beam radiation therapy and concurrent chemotherapy, which was preceded by induction chemotherapy in a subgroup of patients. A total radiation dose of 45 or 50.4 Gy was delivered in daily fractions of 1.8 Gy using three-dimensional computed tomography-based treatment planning (3D-CT), intensity modulated radiation therapy (IMRT), or proton therapy. Chemotherapy generally consisted of a fluoropyrimidine (i.v. or oral) with either a platinum compound or a taxane. After completion of CRT, at the discretion of the treating surgeon either a transthoracic (Ivor-Lewis), transhiatal, total (three-field technique), or minimally invasive esophagectomy was performed with curative intent.

Image Acquisition

Integrated ^{18}F-FDG PET/computed tomography (CT) scans were performed on a dedicated PET/CT system (Discovery RX, ST, STE, or HR; GE Medical Systems, Milwaukee [WI], USA). Patients were instructed to fast for at least 6 hours before ^{18}F-FDG PET and a glucose level within the normal range (80-120 mg/dl) was confirmed. Before ^{18}F-FDG PET, a CT without contrast enhancement was acquired (120 kV peaks, 300 mA, 0.5 seconds rotation, pitch of 1.375, slice thickness 3.75mm, and slice interval 3.27 mm) for attenuation correction purposes. ^{18}F-FDG PET scans were acquired 60-90 minutes after administration of ^{18}F-FDG with a dose of 555-740 MBq, in either two-dimensional (2D) or three-dimensional (3D) acquisition mode at 3-5 minutes per bed position. All images were composed of 128 x 128 pixels with voxel dimensions of 5.47 x 5.47 x 3.27 mm. Images were reconstructed using either attenuation-weighted ordered-subset expectation maximization in 2D (OSEM2D) or iterative reconstruction in 3D (IR3D) images. Specifically, OSEM2D was used in 22 (20%) of 111 patients scanned on Discovery RX, 18 (22%) of 81 patients on Discovery STE, all (100%) of 16 patients on Discovery ST, and all (100%) of 9 patients on Discovery HR. In all cases two iterations, 20-21 subsets, and a 6 mm post-processing Gaussian blurring filter were used.

Supplemental Appendix B. Rationale for Tumor Delineation Method

In the current study a semi-automatic gradient-based delineation method was used that has recently been validated in a multi-observer study reporting superior accuracy, consistency and robustness for target volume contouring compared with manual and threshold methods (*1*). For the purpose of this study we have deliberately chosen not to use thresholding or manual delineation techniques. Although thresholding techniques are most frequently used due to their simple implementation and high efficiency, limitations include difficulty in decision-making for threshold and high sensitivity to tumor heterogeneity, motion artifacts, noise and contrast variations; leading to disappointing results for small, heterogeneous or non-spherical tumors (*2,3*). Manual delineation is simple to apply, but besides time-consuming it is susceptible to window-level settings, suffers from intra- and inter-observer variability and depends on experience of the reader (*2*).

Supplemental Appendix C. Detailed Description of Predictor Selection and Model Building

Clinical Parameters (Model 1)

The potential clinical predictors of pathCR determined at baseline that were included in the analysis were gender, age, body mass index (BMI), co-morbidities (including hypertension, cardiac co-morbidity, diabetes mellitus, and chronic obstructive pulmonary disorder), smoking, Karnofsky performance status, year of patient accrual, tumor location, tumor length based on pre-treatment endoscopic ultrasound (EUS), histologic differentiation grade, signet ring cell adenocarcinoma (yes vs. no), clinical T-stage, and clinical N-stage. TNM-staging was performed in accordance with the seventh edition of the American Joint Committee on Cancer staging manual (*4*). Treatment-related clinical predictors included in the analysis were induction chemotherapy (yes vs. no), total radiation dose, radiation treatment modality, chemotherapy regimen, time interval between completion of CRT and surgery, and the result of post-CRT endoscopic biopsy.

Subjective Assessment ^{18}F-FDG PET (Model 2)

Clinical complete response based on subjective assessment of post-chemoradiation ^{18}F-FDG PET scans by experienced nuclear medicine physicians was defined as having a physiologic level of SUV_{max} at the original primary tumor site or a higher than normal SUV_{max} with an ^{18}F-FDG uptake distribution following an esophagitis pattern (*5*).

Conventional Quantitative ^{18}F-FDG PET Features (Model 3)

The following four conventional quantitative features were extracted from the VOIs of baseline and post-chemoradiation ^{18}F-FDG PET scans: SUV_{max}, SUV_{mean}, MTV and TLG. The TLG was calculated by multiplying MTV by SUV_{mean} (*6*). In addition, the relative changes (in %) of these parameters between baseline and post-chemoradiation scans were calculated and included in the analysis.

Comprehensive Quantitative ^{18}F-FDG PET Features (Model 4)

Various additional features that characterize global, regional and local ^{18}F-FDG uptake intensity distribution and geometry of each tumor were computed from the VOIs of both baseline and post-chemoradiation ^{18}F-FDG PET scans as potential predictors. Also, for each feature the relative change (in %) between the two scans was calculated and included in the analysis.

First-order features were examined to describe global texture related to the SUV frequency distribution (i.e. SUV histogram) within the VOI, and included SUV_{peak} (defined as the average intensity of a 3x3x3 voxel cube centered at the SUV_{max} (*7*)), standard deviation, cumulative histogram, skewness, and kurtosis, among others (*8*). Second-order features describing local texture were calculated using intensity co-occurrence matrices (ICMs) and included parameters such as entropy, energy, homogeneity, and dissimilarity. ICMs determine how often a pixel of intensity *i* finds itself within a certain relationship to another pixel of intensity *j* (*8*).

Also, both higher-order local and regional texture features were included in the analysis. Higher-order local texture features were calculated from neighborhood intensity difference matrices (NIDMs) reflecting differences between each voxel and its neighboring voxels and included busyness, coarseness, contrast, complexity and texture strength. Regional intensity variations were reflected by higher-order regional texture features calculated from voxel alignment (e.g. run-length statistics). A run is defined as a string of consecutive pixels which have the same intensity along a specific linear orientation. Fine textures tend to contain more short runs with similar intensities, whereas coarse textures contain more long runs with different intensities.

For features using a histogram for calculation, a uniform quantization into 100 bins was used. In previous studies, a quantization of at least 64 bins was shown to provide sufficient texture feature reproducibility and robustness (*9-11*). In those studies, 64 bins were recommended because it allowed for SUV increments of 0.25 in their range of encountered SUVs (~4-20), whereas in the current series with a wider range of encountered SUVs, 100 bins allowed for similar SUV increments. Calculations of the NIDM and ICM features were performed for 3 dimensions (i.e. each reference pixel had 26 neighbors [NIDM] and 13 unique directions [ICM]) (*12*). This method allowed NIDM features to use all adjacent pixel values in calculations (and not only those within the same axial slice) and ICM features to be non-directional (*12*).

Pre-selection Method for Inclusion in Multivariable Analysis

Because many potential predictors were studied in univariable analysis, a standardized pre-selection of variables for multivariable analysis was performed according to the following three rules. First, only parameters with a p-value of ≤0.25 in univariable analysis were pre-selected for multivariable analysis. Second, from highly correlated pairs of quantitative ^{18}F-FDG PET parameters (i.e. Spearman rank correlation coefficient r≥0.6) only the one parameter with the lowest p-value in univariable analysis was pre-selected. This rule was applied because high correlations between most of the ^{18}F-FDG PET features were expected (*9*), resulting in the statistical problem of (multi)collinearity in which unstable estimations of regression coefficients for individual predictors occur (*13,14*). Third, only quantitative ^{18}F-FDG PET features with a sufficient robustness (i.e. ICC of ≥0.7) as determined by test-retest analysis were considered eligible for multivariable analysis.

Model Development

Among pre-selected variables for multivariable modeling, no missing values were found. Four multivariable logistic regression models were constructed to study the incremental value of subjective and quantitative assessment of ^{18}F-FDG PET for the prediction of pathCR beyond clinical predictors. Initially, a prediction model with pre-selected clinical parameters only was constructed using stepwise backward elimination of the least significant parameters associated with pathCR based on the Akaike's Information Criterion (AIC) (*15*) (*Model 1*). Subsequently, the final predictors of the clinical model were forced into a second model incorporating the 'subjective assessment of post-chemoradiation ^{18}F-FDG PET' parameter to study the incremental value of this assessment beyond clinical predictors (*Model 2*). Similarly, Model 2 was forced into a third model incorporating conventional quantitative ^{18}F-FDG PET features (*Model 3*), and remaining parameters in Model 3 were forced into a fourth model incorporating comprehensive ^{18}F-FDG PET texture and geometry features (*Model 4*). The AIC-based stepwise backward elimination was repeated for each model with a forced entry (e.g. prohibited elimination) of the parameters of the previous model.

Model Performance and Validation

Model discrimination and calibration results were evaluated for all four models. Discrimination refers to the model's ability to distinguish between patients with pathCR and patients with residual cancer and was assessed by *c*-indexes (*15*). Calibration refers to the agreement between the predicted probability of pathCR and the observed incidences and was evaluated by visual inspection of model calibration plots (*15*). Internal validation using the bootstrap method with 1000 repetitions was carried out to provide insight into potential over-fitting and optimism in model performance. The entire AIC-based backward selection process was repeated in every bootstrap sample to additionally account for the influences (e.g. bias) of the predictor selection steps. Bootstrapping allowed for calculation of bias-corrected *c*-indexes of the four models, and provided shrinkage factors that were used to adjust the estimated regression coefficients in the final four models for over-fitting and miscalibration (*13*).

Clinical Benefit

Diagnostic models such as built in the current study are typically evaluated only with measures of accuracy (e.g. ROC-curve analysis) that do not address clinical consequences. Therefore, a method called decision-curve analysis was developed for evaluating and comparing prediction models that incorporates clinical consequences and requires only the dataset on which the models are tested (*16*). This method assumes that the threshold probability of a certain outcome (e.g. pathCR) at which a patient would opt for a change in treatment (e.g. omission of surgery) weighs the relative harms of a false-positive and a false-negative prediction. This theoretical relationship is then used to derive the “net benefit” of the prediction model across different threshold probabilities. For calculation of the net benefit, the proportion of all patients who are false-positive (e.g. incorrectly classified by the model as complete responder) is subtracted from the proportion who are true-positive (e.g. correctly classified by the model as complete responder), weighted by the relative harm of a false-positive and a false-negative result (i.e. the threshold probability). The “decision curve” is acquired by plotting the net benefit against the threshold probability. As such, the decision-curve analysis identifies the range of threshold probabilities in which a model is of value, the magnitude of the benefit, and which of several models is superior.

References Supplemental Data

(1) Werner-Wasik M, Nelson AD, Choi W, et al. What is the best way to contour lung tumors on PET scans: multiobserver validation of a gradient-based method using a NSCLC digital PET phantom. *Int J Radiat Oncol Biol Phys*. 2012;82:1164-1171.

(2) Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *Eur J Nucl Med Mol Imaging*. 2010;37:2165-2187.

(3) Lee JA. Segmentation of positron emission tomography images: some recommendations for target delineation in radiation oncology. *Radiother Oncol*. 2010;96:302-307.

(4) Edge S, Byrd D, Compton C, editors. *AJCC Cancer Staging Manual*, *7th ed*. New York, NY: Springer; 2010:103-115.

(5) Suzuki A, Xiao L, Hayashi Y, et al. Prognostic significance of baseline positron emission tomography and importance of clinical complete response in patients with esophageal or gastroesophageal junction cancer treated with definitive chemoradiotherapy. *Cancer.* 2011;117:4823-4833.

(6) Roedl JB, Colen RR, Holalkere NS, Fischman AJ, Choi NC, Blake MA. Adenocarcinomas of the esophagus: response to chemoradiotherapy is associated with decrease of metabolic tumor volume as measured on PET-CT: comparison to histopathologic and clinical response evaluation. *Radiother Oncol.* 2008;89:278-286.

(7) Tan S, Kligerman S, Chen W, et al. Spatial-temporal [^{18}F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys.* 2013;85:1375-1382.

(8) Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in ^{18}F-FDG PET/CT imaging by texture analysis. *Eur J Nucl Med Mol Imaging*. 2013;40:133-140.

(9) Hatt M, Majdoub M, Vallieres M, et al. ^{18}F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med.* 2015;56:38-44.

(10) Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in ^{18}F-FDG PET. *J Nucl Med*. 2012;53:693-700.

(11) Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging*. 2013;40:1662-1671.

(12) Fried DV, Tucker SL, Zhou S, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *Int J Radiat Oncol Biol Phys*. 2014;90:834-842.

(13) Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: i. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98:683-690.

(14) Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis Or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162:W1-73.

(15) Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating. 1st ed.* New York, NY: Springer-Verlag; 2009.

(16) Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26:565-574.

SUPPLEMENTAL TABLE 1 Test-retest intraclass correlation coefficients and univariable analysis for predicting pathologic complete response for first- and second-order ^{18}F-FDG PET texture features.

Group	Feature	ICC	Baseline		Post-chemoradiation		Relative difference	
		(95% CI)	OR (95% CI)	*p* value	OR (95% CI)	*p* value	OR (95% CI)	*p* value
Texture: First-order features (*Global*)	SUV$_{peak}$ (*log*)	0.91 (0.45-0.98)	0.62 (0.36-1.06)	0.088	0.24 (0.09-0.59)	0.003*	1.00 (0.99-1.01)	0.990
	Median SUV (*log*)	0.84 (0.04-0.97)	0.59 (0.31-1.12)	0.111	0.70 (0.23-2.08)	0.523	1.01 (1.00-1.02)	0.127
	Minimum SUV (*log*)	0.69 (0.00-0.95)	0.72 (0.33-1.54)	0.399	1.65 (0.52-5.23)	0.389	1.01 (1.00-1.02)	0.070
	Standard deviation [SD] (*log*)	0.86 (0.16-0.98)	0.73 (0.47-1.11)	0.146	0.41 (0.22-0.75)	0.004*	1.00 (0.99-1.01)	0.644
	Variance (*log*)	0.76 (0.00-0.96)	0.85 (0.69-1.06)	0.146	0.64 (0.47-0.86)	0.004*	1.00 (0.99-1.01)	0.623
	Range (*log*)	0.86 (0.17-0.98)	0.70 (0.45-1.07)	0.100	0.34 (0.19-0.56)	<0.001*	0.99 (0.98-1.00)	0.236
	Interquartile range (*log)*	0.73 (0.00-0.95)	0.78 (0.52-1.16)	0.223	0.47 (0.26-0.82)	0.009*	1.00 (0.99-1.01)	0.613
	Cumulative histogram	0.85 (0.11-0.97)	1.02 (0.99-1.05)	0.236	1.09 (1.05-1.14)	<0.001	1.01 (1.00-1.02)	0.078
	Root mean square (*log*)	0.85 (0.11-0.97)	0.61 (0.32-1.11)	0.113	0.61 (0.20-1.75)	0.370	1.01 (1.00-1.02)	0.164
	Skewness	0.79 (0.00-0.96)	0.91 (0.40-2.04)	0.824	0.43 (0.21-0.85)	0.017*	0.99 (0.98-1.00)	0.043*
	Kurtosis (*log*)	0.40 (0.00-0.90)	0.82 (0.28-2.34)	0.718	0.12 (0.04-0.35)	<0.001*	0.99 (0.98-1.00)	0.024*
	Energy (*log*)	0.83 (0.00-0.97)	0.86 (0.70-1.05)	0.133	0.52 (0.38-0.70)	<0.001*	1.00 (0.99-1.00)	0.531
	Entropy	0.93 (0.56-0.99)	0.77 (0.55-1.06)	0.115	0.44 (0.25-0.73)	0.002*	1.00 (0.98-1.00)	0.132
	Uniformity (*log*)	0.95 (0.68-0.99)	1.44 (0.91-2.31)	0.122	3.04 (1.44-6.76)	0.005*	1.00 (1.00-1.00)	0.562
Texture: Second-order features (*Local*)	Local maximum entropy (*log*)	0.86 (0.19-0.98)	0.13 (0.01-1.12)	0.065	0.03 (0.00-0.20)	0.001*	0.98 (0.96-1.01)	0.139
	Local mean entropy (*log*)	0.88 (0.30-0.98)	0.14 (0.02-1.17)	0.071	0.04 (0.00-0.28)	0.002*	0.99 (0.96-1.01)	0.184
	Local median entropy (*log*)	0.89 (0.34-0.98)	0.16 (0.02-1.32)	0.090	0.05 (0.01-0.34)	0.004*	0.99 (0.96-1.01)	0.211
	Local minimum entropy (*log*)	0.82 (0.00-0.97)	0.30 (0.04-2.19)	0.232	0.10 (0.02-0.55)	0.010*	0.99 (0.97-1.01)	0.211
	Local entropy SD (*log*)	0.81 (0.00-0.97)	0.61 (0.20-1.76)	0.361	1.10 (0.45-2.72)	0.835	1.01 (1.00-1.01)	0.164
	Local maximum range (*log*)	0.86 (0.16-0.98)	0.89 (0.43-1.09)	0.116	0.37 (0.19-0.67)	0.002*	1.00 (0.99-1.01)	0.467
	Local mean range (*log*)	0.89 (0.34-0.98)	0.69 (0.43-1.10)	0.122	0.44 (0.23-0.82)	0.012*	1.00 (0.99-1.01)	0.838
	Local median range (*log*)	0.88 (0.28-0.98)	0.73 (0.46-1.16)	0.189	0.49 (0.25-0.91)	0.027*	1.00 (0.99-1.01)	0.890
	Local minimum range (*log*)	0.87 (0.24-0.98)	0.61 (0.35-1.06)	0.082	0.64 (0.32-1.26)	0.197	1.00 (1.00-1.01)	0.230
	Local range SD (*log*)	0.74 (0.00-0.96)	0.76 (0.50-1.14)	0.185	0.39 (0.24-0.60)	<0.001*	0.99 (0.98-1.00)	0.087
	Local maximum SD (*log*)	0.85 (0.15-0.98)	0.75 (0.48-1.17)	0.206	0.40 (0.21-0.75)	0.006*	1.00 (0.99-1.01)	0.579
	Local mean SD (*log*)	0.87 (0.25-0.98)	0.73 (0.46-1.13)	0.162	0.52 (0.26-1.01)	0.057	1.00 (0.99-1.01)	0.789
	Local median SD (*log*)	0.86 (0.16-0.98)	0.76 (0.48-1.17)	0.211	0.58 (0.29-1.10)	0.101	1.00 (0.99-1.01)	0.768
	Local minimum SD (*log*)	0.84 (0.09-0.97)	0.63 (0.35-1.10)	0.109	0.80 (0.40-1.60)	0.529	1.01 (1.00-1.01)	0.082
	Local SD SD (*log*)	0.77 (0.00-0.96)	0.82 (0.55-1.23)	0.334	0.39 (0.23-0.63)	<0.001*	0.99 (0.99-1.00)	0.100
	Mean absolute deviation (*log*)	0.84 (0.05-0.97)	0.74 (0.48-1.12)	0.160	0.42 (0.23-0.75)	0.004*	1.00 (0.99-1.01)	0.618
	Median absolute deviation (*log*)	0.74 (0.00-0.96)	0.78 (0.52-1.16)	0.228	0.44 (0.25-0.74)	0.002*	1.00 (0.99-1.01)	0.522
	Autocorrelation (*log*)	0.83 (0.03-0.97)	0.76 (0.54-1.06)	0.115	0.70 (0.37-1.28)	0.259	1.00 (1.00-1.01)	0.171
	Cluster prominence (*log*)	0.49 (0.00-0.91)	0.92 (0.83-1.03)	0.139	0.71 (0.59-0.85)	<0.001*	1.00 (0.99-1.00)	0.637
	Cluster shade (*log*)	0.70 (0.00-0.95)	0.31 (0.09-1.10)	0.070	0.28 (0.07-0.77)	0.037*	1.00 (1.00-1.00)	0.660
	Cluster tendency (*log*)	0.74 (0.00-0.96)	0.86 (0.69-1.06)	0.153	0.47 (0.30-0.71)	0.001*	1.00 (0.99-1.00)	0.475
	ICM contrast (*log*)	0.73 (0.00-0.95)	0.83 (0.66-1.04)	0.106	0.57 (0.37-0.86)	0.009*	1.00 (0.99-1.00)	0.866
	Correlation (*log*)	0.59 (0.00-0.93)	1.26 (0.87-1.91)	0.256	0.11 (0.02-0.46)	0.003*	1.00 (1.00-1.00)	0.130
	Difference entropy (*log*)	0.87 (0.25-0.98)	0.43 (0.16-1.17)	0.098	0.15 (0.05-0.37)	<0.001*	0.99 (0.98-1.00)	0.049*
	Dissimilarity (*log*)	0.81 (0.00-0.97)	0.69 (0.44-1.07)	0.096	0.37 (0.16-0.81)	0.015*	1.00 (0.99-1.01)	0.829
	ICM Energy (*log*)	0.95 (0.71-0.99)	1.28 (0.99-1.66)	0.060	25.9 (5.10-155)	<0.001*	1.00 (1.00-1.01)	0.095
	ICM Entropy	0.93 (0.56-0.99)	0.85 (0.70-1.01)	0.070	0.59 (0.45-0.76)	<0.001*	0.98 (0.97-0.99)	0.016*
	Homogeneity 1	0.90 (0.41-0.98)	7.68 (0.77-77.6)	0.081	9.37 (0.94-99.4)	0.059	1.00 (1.00-1.00)	0.878
	Homogeneity 2	0.91 (0.49-0.99)	5.96 (0.77-46.3)	0.086	6.95 (1.01-50.3)	0.051	1.00 (1.00-1.00)	0.883
	Informational measure correlation 1 (*log*)	0.63 (0.00-0.94)	0.99 (0.23-4.87)	0.984	0.02 (0.00-0.20)	0.004*	1.00 (1.00-1.01)	0.015*
	Informational measure correlation 2	0.79 (0.00-0.96)	0.66 (0.12-3.70)	0.631	3.12 (0.69-14.3)	0.138	1.01 (1.00-1.01)	0.149
	Inverse difference moment normalized (*log*)	0.73 (0.00-0.95)	1.34 (0.45-7.03)	0.545	97.9 (1.86-2351)	0.065	0.82 (0.45-1.30)	0.441
	Inverse difference normalized (log)	0.82 (0.00-0.97)	1.65 (0.72-5.27)	0.320	18.4 (1.44-415)	0.048*	0.95 (0.85-1.06)	0.402
	Inverse variance (*log*)	0.88 (0.29-0.98)	1.62 (0.88-3.08)	0.133	1.70 (0.23-13.9)	0.609	1.00 (1.00-1.00)	0.468
	Maximum probability (*log*)	0.91 (0.50-0.99)	1.34 (1.01-1.81)	0.046*	1.90 (1.26-2.94)	0.003*	1.00 (1.00-1.00)	0.243
	Sum average (*log*)	0.85 (0.11-0.97)	0.58 (0.29-1.12)	0.112	0.50 (0.14-1.66)	0.267	1.01 (1.00-1.02)	0.200
	Sum entropy	0.93 (0.58-0.99)	0.79 (0.58-1.07)	0.124	0.40 (0.26-0.61)	<0.001*	0.98 (0.96-0.99)	0.004*
	Sum variance (*log*)	0.83 (0.01-0.97)	0.78 (0.57-1.05)	0.108	0.92 (0.51-1.64)	0.792	1.01 (1.00-1.01)	0.054
	ICM Variance (*log*)	0.74 (0.00-0.96)	0.86 (0.69-1.06)	0.153	0.47 (0.30-0.71)	0.001*	1.00 (0.99-1.00)	0.475

SUPPLEMENTAL TABLE 2 Test-retest intraclass correlation coefficients and univariable analysis for predicting pathologic complete response for higher-order ^{18}F-FDG PET texture features and geometry features.

Group	Feature	ICC	Baseline		Post-chemoradiation		Relative difference	
			OR (95% CI)	*p* value	OR (95% CI)	*p* value	OR (95% CI)	*p* value
Texture: Higher-order features (*Local*)	Busyness (*log*)	0.88 (0.29-0.98)	1.33 (0.93-1.93)	0.117	2.12 (1.53-3.02)	<0.001*	1.00 (1.00-1.01)	0.005*
	Coarseness	0.52 (0.00-0.92)	3.54 (1.05-12.6)	0.044*	0.21 (0.04-1.05)	0.069	1.00 (0.99-1.00)	0.051
	Complexity (*log*)	0.68 (0.00-0.95)	0.83 (0.67-1.01)	0.073	0.59 (0.42-0.81)	0.002*	1.00 (1.00-1.00)	0.961
	Contrast	0.63 (0.00-0.94)	9.68 (0.06-2159)	0.360	1.20 (0.05-13.1)	0.874	1.00 (1.00-1.01)	0.034*
	Texture strength (*log*)	0.74 (0.00-0.96)	0.79 (0.53-1.14)	0.221	0.44 (0.11-1.37)	0.199	1.00 (1.00-1.00)	0.572
Texture: Higher-order features (*Regional*)	Intensity non-uniformity	0.89 (0.34-0.98)	1.01 (0.99-1.03)	0.434	0.94 (0.89-0.98)	0.006*	1.00 (0.99-1.00)	0.101
	Run length non-uniformity (*log*)	0.97 (0.81-0.99)	0.81 (0.57-1.14)	0.223	0.43 (0.29-0.61)	<0.001*	0.99 (0.99-1.00)	0.136
	Run percentage (*log*)	0.89 (0.38-0.98)	0.47 (0.18-1.09)	0.087	1.70 (0.88-4.07)	0.181	1.03 (1.01-1.06)	0.021*
	High intensity run emphasis (*log*)	0.84 (0.04-0.97)	0.77 (0.54-1.07)	0.121	0.75 (0.40-1.37)	0.356	1.01 (1.00-1.01)	0.127
	Low intensity run emphasis (*log*)	0.84 (0.06-0.97)	1.34 (0.92-1.98)	0.135	1.22 (0.65-2.35)	0.544	1.00 (1.00-1.00)	0.976
	Long run emphasis	0.90 (0.42-0.98)	2.13 (0.92-5.22)	0.079	0.52 (0.23-1.11)	0.110	0.99 (0.98-1.00)	0.025*
	Short run emphasis (*log*)	0.90 (0.43-0.98)	0.33 (0.08-1.26)	0.102	1.63 (0.47-6.63)	0.462	1.03 (1.00-1.06)	0.067
	Long run high intensity emphasis (*log*)	0.81 (0.00-0.97)	0.75 (0.50-1.11)	0.159	0.43 (0.19-0.92)	0.034*	1.00 (1.00-1.01)	0.455
	Short run high intensity emphasis (*log*)	0.84 (0.07-0.97)	0.77 (0.56-1.06)	0.116	0.84 (0.48-1.43)	0.514	1.00 (1.00-1.01)	0.106
	Long run low intensity emphasis (*log*)	0.86 (0.16-0.98)	1.31 (0.96-1.82)	0.095	0.97 (0.61-1.56)	0.900	1.00 (1.00-1.00)	0.792
	Short run low intensity emphasis (*log)*	0.83 (0.00-0.97)	1.34 (0.90-2.03)	0.153	1.33 (0.67-2.74)	0.432	1.00 (1.00-1.00)	0.792
Geometry (*Size and shape*)	Maximum 3D diameter (*log*)	0.98 (0.85-1.00)	0.58 (0.24-1.39)	0.222	0.16 (0.06-0.39)	<0.001*	0.99 (0.98-1.00)	0.127
	Compactness (*log*)	0.99 (0.97-1.00)	0.75 (0.37-1.53)	0.433	0.13 (0.06-0.29)	<0.001*	0.98 (0.97-0.99)	0.009*
	Convex (*log*)	0.83 (0.00-0.97)	0.42 (0.15-1.17)	0.093	1.36 (0.35-8.23)	0.690	1.15 (1.00-1.33)	0.051
	Convex hull volume 2D (*log*)	0.99 (0.99-1.00)	0.86 (0.58-1.26)	0.436	0.34 (0.21-0.52)	<0.001*	1.00 (0.99-1.00)	0.177
	Convex hull volume 3D (*log*)	0.99 (0.99-1.00)	0.85 (0.60-1.22)	0.379	0.43 (0.30-0.60)	<0.001*	1.00 (0.99-1.00)	0.242
	Mean breadth (*log*)	0.93 (0.61-0.99)	0.59 (0.21-1.65)	0.318	0.07 (0.02-0.21)	<0.001*	0.98 (0.97-0.99)	0.013*
	Orientation (*log*)	0.86 (0.18-0.98)	1.23 (0.83-1.91)	0.316	0.94 (0.73-1.23)	0.622	1.00 (0.99-1.00)	0.321
	Roundness (*log*)	0.84 (0.05-0.97)	0.50 (0.11-2.24)	0.365	0.33 (0.10-1.05)	0.062	1.00 (0.99-1.01)	0.386
	Spherical disproportion	0.69 (0.00-0.95)	0.62 (0.03-9.94)	0.743	5.27 (0.46-61.1)	0.181	1.02 (0.99-1.05)	0.194
	Sphericity (*log*)	0.67 (0.00-0.94)	1.98 (0.03-152)	0.750	0.07 (0.00-3.08)	0.171	0.98 (0.95-1.01)	0.234
	Surface area (*log*)	0.99 (0.97-1.00)	0.77 (0.42-1.39)	0.386	0.22 (0.11-0.41)	<0.001*	0.99 (0.98-1.00)	0.070
	Surface area density	0.97 (0.80-0.99)	1.44 (0.64-3.26)	0.376	3.99 (2.28-7.40)	<0.001*	1.01 (1.01-1.02)	<0.001*

SUPPLEMENTAL TABLE 3 Logistic regression formulas of the four prediction models.

Model no.	Shrinkage factor	Regression formula
1	0.713	$\log(p/1\text{-}p) = 0.647 + \mathbf{0.713\ (}\ \text{-}0.741{*}\log(\mathit{EUSlength}) - 0.796{*}\mathit{cTstage} + 0.894{*}\mathit{IndChTx} - 1.20{*}\mathit{EndoBiopsy}\mathbf{)}$
2	0.839	$\log(p/1\text{-}p) = 1.53 + \mathbf{0.839\ (}\ \text{-}0.700{*}\log(\mathit{EUSlength}) - 0.943{*}\mathit{cTstage} + 0.877{*}\mathit{IndChTx} - 1.14{*}\mathit{EndoBiopsy} - 1.22{*}\mathit{SubjectivePET}\mathbf{)}$
3	0.798	$\log(p/1\text{-}p) = 2.34 + \mathbf{0.798\ (}\ \text{-}0.598{*}\log(\mathit{EUSlength}) - 0.643{*}\mathit{cTstage} + 0.815{*}\mathit{IndChTx} - 1.15{*}\mathit{EndoBiopsy} - 0.791{*}\mathit{SubjectivePET} - 0.565{*}\log(\mathit{pTLG})\mathbf{)}$
4	0.743	$\log(p/1\text{-}p) = 16.5 + \mathbf{0.743\ (}\ \text{-}0.777{*}\log(\mathit{EUSlength}) - 0.618{*}\mathit{cTstage} + 1.03{*}\mathit{IndChTx} - 1.19{*}\mathit{EndoBiopsy} - 0.663{*}\mathit{SubjectivePET} - 0.280{*}\log(\mathit{pTLG}) - 1.68{*}\log(\mathit{bClusterShade} + 1500) - 0.029{*}\mathit{dICMEntropy} + 0.063{*}\mathit{dRunPercentage} - 2.27{*}\log(10{*}\mathit{pRoundness})\mathbf{)}$

p: Probability of pathologic complete response. *EUSlength*: EUS-based tumor length (in cm) at baseline. *cTstage*: Clinical T-stage (0=cT2, 1=cT3). *IndChTx*: Induction chemotherapy (no=0, yes=1). *EndoBiopsy*: Result of post-chemoradiation endoscopic biopsy (0=negative, 1=positive). *SubjectivePET*: Subjective assessment of response on post-chemoradiation ^{18}F-FDG PET (0=complete response, 1=residual cancer). *pTLG*: Post-chemoradiation Total lesion glycolysis. *bClusterShade*: Baseline Cluster shade. *dICMEntropy*: Relative difference (in %) of ICM Entropy between baseline and post-chemoradiation ^{18}F-FDG PET. *dRunPercentage*: Relative difference (in %) of Run percentage between baseline and post-chemoradiation ^{18}F-FDG PET. *pRoundness*: Post-chemoradiation Roundness.

Note. **Bold** numbers represent shrinkage factors (provided by bootstrapping) used to adjust the estimated regression coefficients for over-fitting and miscalibration.