# Prediction of Response to Neoadjuvant Chemotherapy and Radiation Therapy with Baseline and Restaging F-18-FDG PET Imaging Biomarkers in Patients with Esophageal Cancer

Beukinga, Roelof J; Hulshoff, Jan Binne; Mul, Véronique E M; Noordzij, Walter; Kats-Ugurlu, Gursah; Slart, Riemer H J A; Plukker, John T M

*Published in:*
Radiology

*DOI:*
10.1148/radiol.2018172229

# Prediction of Response to Neoadjuvant Chemotherapy and Radiation Therapy with Baseline and Restaging $^{18}$F-FDG PET Imaging Biomarkers in Patients with Esophageal Cancer[1]

Roelof J. Beukinga, MSc
Jan Binne Hulshoff, MD
Véronique E. M. Mul, MD
Walter Noordzij, MD, PhD
Gursah Kats-Ugurlu, MD
Riemer H. J. A. Slart, MD, PhD
John T. M. Plukker, MD, PhD

**Purpose:** To assess the value of baseline and restaging fluorine 18 ($^{18}$F) fluorodeoxyglucose (FDG) positron emission tomography (PET) radiomics in predicting pathologic complete response to neoadjuvant chemotherapy and radiation therapy (NCRT) in patients with locally advanced esophageal cancer.

**Materials and Methods:** In this retrospective study, 73 patients with histologic analysis–confirmed T1/N1–3/M0 or T2–4a/N0–3/M0 esophageal cancer were treated with NCRT followed by surgery (Chemoradiotherapy for Esophageal Cancer followed by Surgery Study regimen) between October 2014 and August 2017. Clinical variables and radiomic features from baseline and restaging $^{18}$F-FDG PET were selected by univariable logistic regression and least absolute shrinkage and selection operator. The selected variables were used to fit a multivariable logistic regression model, which was internally validated by using bootstrap resampling with 20000 replicates. The performance of this model was compared with reference prediction models composed of maximum standardized uptake value metrics, clinical variables, and maximum standardized uptake value at baseline NCRT radiomic features. Outcome was defined as complete versus incomplete pathologic response (tumor regression grade 1 vs 2–5 according to the Mandard classification).

**Results:** Pathologic response was complete in 16 patients (21.9%) and incomplete in 57 patients (78.1%). A prediction model combining clinical T-stage and restaging NCRT (post-NCRT) joint maximum (quantifying image orderliness) yielded an optimism-corrected area under the receiver operating characteristics curve of 0.81. Post-NCRT joint maximum was replaceable with five other redundant post-NCRT radiomic features that provided equal model performance. All reference prediction models exhibited substantially lower discriminatory accuracy.

**Conclusion:** The combination of clinical T-staging and quantitative assessment of post-NCRT $^{18}$F-FDG PET orderliness (joint maximum) provided high discriminatory accuracy in predicting pathologic complete response in patients with esophageal cancer.



*Online supplemental material is available for this article.*

[1] From the Departments of Surgical Oncology (R.J.B., J.B.H., J.T.M.P.), Nuclear Medicine and Molecular Imaging (R.J.B., W.N., R.H.J.A.S.), Radiology (J.B.H.), Radiation Oncology (V.E.M.M.), and Pathology (G.K.U.), University of Groningen, University Medical Center Groningen, Hanzeplein 1, Groningen 9713 GZ, the Netherlands; and Department of Biomedical Photonic Imaging, University of Twente, Enschede, the Netherlands (R.J.B., R.H.J.A.S.). Received October 4, 2017; revision requested November 10; revision received December 20; accepted December 21. **Address correspondence to** R.J.B. (e-mail: *r.j.beukinga@umcg.nl*).

Neoadjuvant chemotherapy and radiation therapy (NCRT) followed by esophagectomy is the common standard treatment of resectable locally advanced esophageal cancer (1,2). Pathologic complete response after NCRT is generally achieved in 25%–42% of patients with esophageal cancer and is accompanied with a lower rate of recurrence and longer survival (1,3–6). Some patients without gross residual tumor at restaging evaluation (ie, clinical complete response) may benefit from a so-called wait-and-see policy. However, improving the identification of complete responders is required to omit surgery after NCRT on an individualized basis. The current standard method to predict response to NCRT is the semiquantitative measurement of temporal change in maximum standardized uptake value between baseline NCRT (pre-NCRT) and restaging NCRT (post-NCRT) fluorine 18 ($^{18}$F) fluorodeoxyglucose (FDG) positron emission tomography (PET) (7), however, this value yields an insufficient sensitivity and specificity of 67% and 68%, respectively (8). One of the reasons for this inadequate performance may be that maximum standardized uptake value is a single-voxel representation that is susceptible to noise artifacts (9). Moreover, maximum standardized uptake value ignores the intratumoral $^{18}$F-FDG spatial distribution and does not represent the overall tumor burden. Intratumoral heterogeneity is present in nearly all esophageal tumors and it has been hypothesized (10) that high intratumoral $^{18}$F-FDG uptake heterogeneity before NCRT, usually because of hypoxia, is associated with an impaired tumor response to NCRT.

Radiomics extracts a large number of quantitative imaging features from medical images and may improve image interpretation by acquiring in vivo tumor information. In earlier studies (11–17), $^{18}$F-FDG PET radiomic features that quantified geometric, intensity, and textural ($^{18}$F-FDG spatial distribution) characteristics of tumors exhibited higher diagnostic accuracies than the maximum standardized uptake value for prediction of response in patients with esophageal cancer. Acquiring both pre-NCRT and post-NCRT $^{18}$F-FDG PET images enables close follow-up of tumor response during treatment. Changes in radiomic features during treatment may reflect changes in intratumoral $^{18}$F-FDG heterogeneity and tumor phenotype. Changes in radiomic features have shown (11–13,16,18) promising results for prediction of response in patients with esophageal cancer.

The aim of our study was to assess the value of radiomic features extracted from baseline and restaging $^{18}$F-FDG PET scans in prediction of pathologic complete response to NCRT in patients with locally advanced esophageal cancer.

### Implications for Patient Care

- Primary tumor invasion depth and a radiomic feature (quantifying image orderliness) derived from fluorine 18 ($^{18}$F) fluorodeoxyglucose (FDG) PET images yielded good performance to predict pathologic complete versus incomplete response to neoadjuvant chemotherapy and radiation therapy (NCRT) in esophageal cancer.
- Six different radiomics parameters derived from $^{18}$F-FDG PET images were interchangeable, providing equally good predictive performance to predict pathologic complete versus incomplete response in combination with depth of esophageal carcinoma invasion.
- A prediction model combining radiomic features and tumor invasion depth may guide the decision on whether surgery could be omitted after NCRT in patients with esophageal cancer.

## Materials and Methods

This retrospective study was conducted according to the national Dutch guidelines for retrospective studies and rules of the local institutional ethical board, and the need to obtain informed consent was waived. Patients were eligible for inclusion if they had locally advanced (T1, N1–3, M0; or T2–4a, N0–3, M0) esophageal cancer and were treated with NCRT (Chemoradiotherapy for Esophageal cancer followed by Surgery Study regimen [known as CROSS]) followed by esophagectomy with a two-field lymph node dissection in the University Medical Center Groningen (Groningen, the Netherlands) between October 2014 and August 2017 (1). Excluded from the analyses were patients with missing data, those with $^{18}$F-FDG PET/computed tomography (CT) scans performed in other medical centers, those with tumors that were not $^{18}$F-FDG avid, and those with distant metastases found before or during surgery. Overall, the following 73 patients (mean age, 64.4 years ± 8.3; age range, 42–83 years) were consecutively included: 58 men (mean age, 64.0 years ± 7.8; age range, 51–81 years) and 15 women (mean age, 65.7 years ± 9.9; age range, 42–83 years). Clinical data were obtained from a prospectively maintained database (Table 1). Twenty-two patients were reported in an earlier article (15). Our study can be considered as a continuation of this earlier article in which only radiomic features derived from baseline $^{18}$F-FDG PET/CT scans were analyzed (15), whereas in this study we analyzed radiomic

**https://doi.org/10.1148/radiol.2018172229**

**Content code:** NM

**Radiology 2018;** 287:983–992

**Abbreviations:**
FDG = fluorodeoxyglucose
NCRT = neoadjuvant chemotherapy and radiation therapy
preNCRT = baseline NCRT
postNCRT = restaging NCRT

**Author contributions:**
Guarantors of integrity of entire study, R.J.B., J.B.H., J.T.M.P.; 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, R.J.B., J.B.H., V.E.M.M., J.T.M.P.; clinical studies, all authors; experimental studies, J.T.M.P.; statistical analysis, R.J.B., J.B.H., J.T.M.P.; and manuscript editing, R.J.B., J.B.H., V.E.M.M., W.N., R.H.J.A.S., J.T.M.P.

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

**Table 1**

**Patient and Tumor Characteristics**

| Characteristics | Patients ($n = 73$) |
|---|---|
| Sex | |
| Male | 58 (79.5) |
| Female | 15 (20.5) |
| Median age (y)* | 63.0 (59.0–69.0) |
| Histologic analysis result | |
| Adenocarcinoma | 65 (89.0) |
| Squamous cell carcinoma | 8 (11.0) |
| Tumor location | |
| Middle esophagus | 9 (12.3) |
| Distal esophagus/GEJ | 64 (87.7) |
| Median tumor length (cm)* | 6.0 (5.0–9.0) |
| Clinical T-stage | |
| T2 | 9 (12.3) |
| T3 | 59 (80.8) |
| T4a | 5 (6.8) |
| Clinical N-stage | |
| N0 | 15 (20.5) |
| N1 | 32 (43.8) |
| N2 | 22 (30.1) |
| N3 | 4 (5.5) |
| No. of chemotherapy cycles | |
| 2 | 1 (1.4) |
| 3 | 1 (1.4) |
| 4 | 12 (16.4) |
| 5 | 59 (80.8) |
| Radicality | |
| R0 | 70 (95.9) |
| R1 | 3 (4.1) |
| Mandard tumor regression grade | |
| 1 | 16 (21.9) |
| 2 | 19 (26.0) |
| 3 | 23 (31.5) |
| 4 | 14 (19.2) |
| 5 | 1 (1.4) |

Note.—Unless otherwise indicated, data are number of patients and data in parentheses are percentages. Radicality refers to histologic complete resection. GEJ = gastroesophageal junction, R0 = microscopically radical resection, R1 = microscopically irradical (ie, histologic noncomplete resection) resection/circumferential resection margin greater than 1 mm.

*Data in parentheses are interquartile range.

features derived from both baseline and restaging $^{18}$F-FDG PET scans.

Staging consisted of 64 multi–detector row CT of the thorax and abdomen, $^{18}$F-FDG PET/CT, and endoscopic ultrasonography with fine-needle aspiration, if indicated. After staging, we discussed all patients in our multidisciplinary upper gastrointestinal tumor board. All patients were treated with NCRT according to the CROSS regimen consisting of carboplatin (2 mg · min · mL$^{-1}$) and paclitaxel (50 mg/m$^2$) in five cycles combined with 41.4 Gy in 23 fractions. All patients were restaged with $^{18}$F-FDG PET/CT approximately 6–8 weeks after NCRT. Surgery consisted of either open or minimally invasive transthoracic esophagectomy combined with a two-field lymph node dissection. Two experienced gastrointestinal pathologists (one of whom was G.K.U., with more than 10 years of experience) determined tumor response to NCRT according to the Mandard tumor regression grade (19), which was considered the standard. This five-point scoring system classifies the percentage of residual vital tumor cells and the degree of NCRT-induced fibrosis. Response was categorized into pathologic complete response (tumor regression grade 1) versus pathologic incomplete response (tumor regression grade 2–5). Because the clinical relevance of categorizing response into pathologic good response (tumor regression grade 1–2) versus poor response (tumor regression grade 3–5) remains unclear, this distinction was not made.

Integrated baseline and restaging $^{18}$F-FDG PET/CT (Biograph mCT-64 PET/CT; Siemens, Knoxville, Tenn) scans were performed. Patients were instructed to fast except for the consumption of water for at least 6 hours before administration of 3 MBq/kg $^{18}$F-FDG. Serum glucose levels were evaluated just before tracer injection. Sixty minutes after tracer injection, continuous breathing low-dose CT images (80–120 kV; 20–35 mAs; and 5-mm section thickness) for viewing anatomic structures and PET images were acquired with the patient positioned in radiation treatment planning position. PET images were obtained with 2–3 minutes per bed position in three-dimensional setting. Images were reconstructed according to the European Association of Nuclear Medicine guidelines (20) by using a time-of-flight iterative reconstruction method (three iterations; 21 subsets; and voxel size, 3.1819 × 3.1819 × 2 mm) with point-spread-function correction. Images were corrected for random coincidences, scatter, and attenuation, and were smoothed with a Gaussian filter of 6.5 mm in full-width at half-maximum.

### Volume-of-Interest Delineation

On the basis of the radiotherapeutic gross tumor volume, which was manually delineated by an expert radiation oncologist (V.E.M.M.) in gastrointestinal malignancies, the volume of interest was defined. The gross tumor volume was rigidly co-registered to the CT component of the baseline and restaging PET/CT images (RTx Workstation 1.0; Mirada Medical, Oxford, England). Because registration errors could occur, the volume of interest was manually corrected after consensus of the collaborating investigators. The post-NCRT delineation included the pre-NCRT localization of the primary tumor, and was adjusted manually to compensate for regression of the tumor size.

### Predictors and Radiomic Feature Extraction

We analyzed the following clinical parameters: sex, age, histologic analysis result, tumor location, tumor length, clinical T-stage, and clinical N-stage. Moreover, we extracted a total of 113 radiomic features from both pre-NCRT and post-NCRT PET images. Software was developed in-house with Matlab 2014b (Mathworks, Natick, Mass) for feature extraction and image processing (21). We extracted 19 morphologic features, two local intensity features, 18 statistical features, and 62 textural features (25 gray-level co-occurrence–based features, 16 gray-level run-length–based features, 16 gray-level size-zone–based features, and five neighborhood gray-tone difference–based features). Among these radiomic features were the following five traditional radiomic features: volume, maximum standardized uptake value, peak standardized uptake value, mean standardized uptake value, and total lesion glycolysis (tumor volume multiplied by mean standardized

$$\Delta\text{-NCRT radiomic feature} = \frac{\text{postNCRT radiomic feature} - \text{preNCRT radiomic feature}}{\text{preNCRT radiomic feature}} \times 100\%,$$

uptake value). For each of these radiomic features, the relative difference between pre-NCRT and post-NCRT $^{18}$F-FDG PET was calculated according to the following equation: where Δ-*NCRT* is the relative difference between pre-NCRT and post-NCRT $^{18}$F-FDG PET. Furthermore, eight bin-to-bin histogram distances and four cross-bin histogram distances were selected, which quantified the perceived similarity in intensity distribution between the pre-NCRT and post-NCRT PET images.

$^{18}$F-FDG uptake was converted to standardized uptake value and was corrected for the serum glucose level (20). Original voxel dimensions were up-sampled to isotropic voxel-dimensions of 2 × 2 × 2 mm by using trilinear spline interpolation. Voxels, enclosed for 50% or greater coverage, were included to the up-sampled volume of interest. Textural features were extracted from discretized image stacks to reduce the continuous-scaled standardized uptake value to a limited number of gray levels and to reduce image noise. Voxels were discretized in 0.5 g/mL increments starting at a minimum of 0 g/mL. Images were analyzed in three dimensions with a connectivity of 26 voxels (13 angular directions and a pixel-to-pixel distance of one). The 13 different gray-level co-occurrence matrices and gray-level run-length matrices along each angular direction were merged into combined matrices before feature extraction.

**Figure 1**

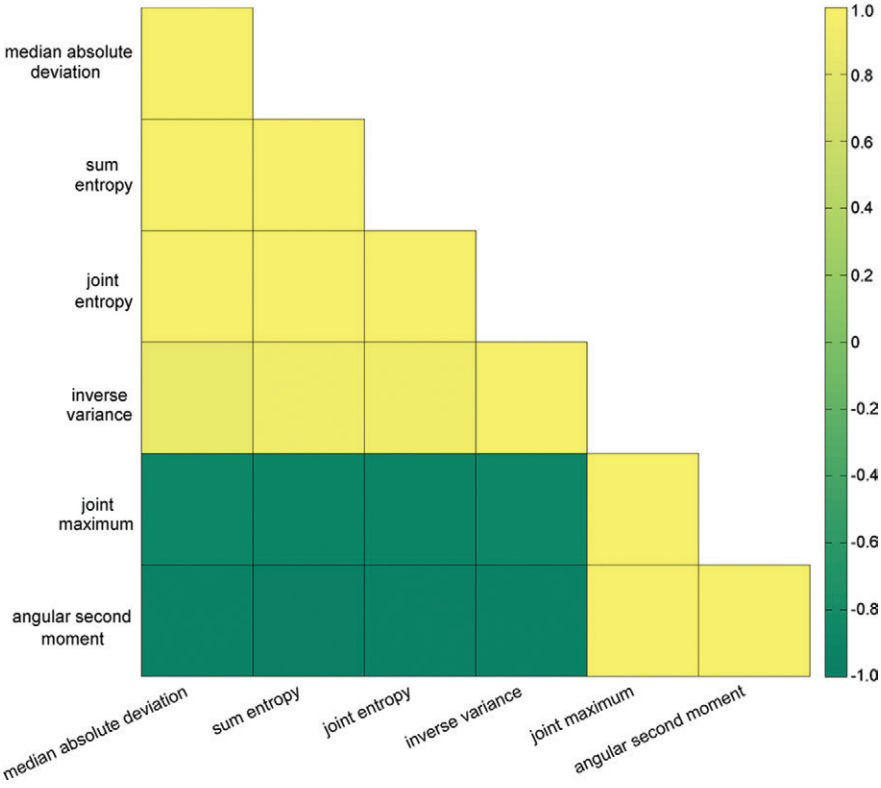



**Figure 1:** Correlation matrix of all radiomic features that positively (Spearman $\rho > 0.9$) or negatively (Spearman $\rho < -0.9$) correlated with restaging NCRT joint maximum. All mentioned radiomic features were extracted from restaging NCRT imaging. The bar represents Spearman correlation.

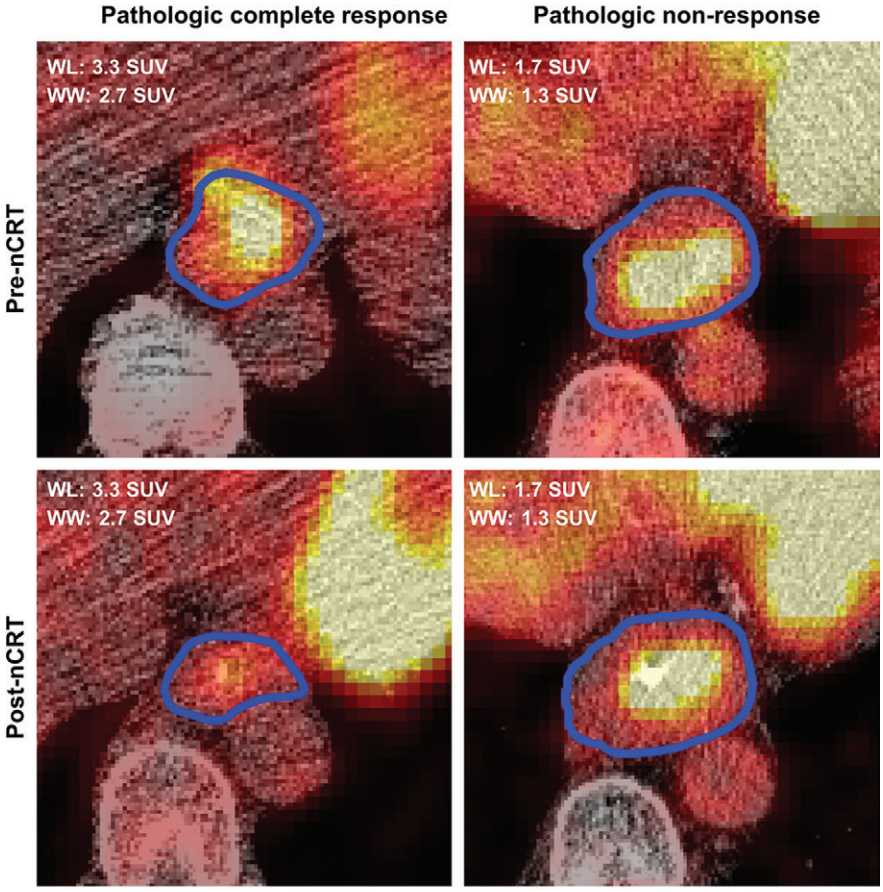
### Statistical Analysis

Statistical analysis was performed with Matlab 2014b (Mathworks) and R 3.2.2 open-source software (Comprehensive R Archive Network, *http://www.r-project.org*). Radiomic features were tested for their robustness to delineation variations. Therefore, two additional segmentations were created by morphologic dilation with two ball-shaped structuring elements with radii of 1 and 2 voxels. The reliability of ratings was measured by the intraclass correlation coefficient. Only radiomic features that had an excellent reliability (intraclass correlation coefficient, >0.75) at both the baseline and the restaging measurements were considered robust. Furthermore, all predictors were tested in a univariable logistic regression with a significance level α of 0.157 according to the Akaike Information Criterion requiring $\chi^2 > 2 \cdot df$, where *df* is degrees of freedom.

Only robust predictors and predictors significant at univariable logistic regression were introduced to a least absolute shrinkage and selection operator for variable selection. The least absolute shrinkage and selection operator shrinks estimated regression coefficients and excludes variables by forcing certain coefficients to become 0 to reduce overfitting (Appendix E1 [online]). Logistic regression models were fitted with the selected variables.

The final model was tested for multicollinearity, as quantified by a variance inflation factor greater than four. The goodness-of-fit of each model was evaluated with the 2-log likelihood, the Akaike information criterion, and the Nagelkerke $R^2$. Each model was quantified in terms of discrimination with the area under the receiver operating characteristic curve and the discrimination slope and quantified in terms of calibration by using the slope and intercept of calibration plots and the Hosmer-Lemeshow test. Because the models are trained and tested on the same dataset,

**Figure 2**




| | Pre-nCRT | Post-nCRT | Δ-nCRT (%) | Pre-nCRT | Post-nCRT | Δ-nCRT (%) |
|---|---|---|---|---|---|---|
| MAD | 0.56 | 0.16 | -71 | 3.56 | 4.31 | 21 |
| JM | 0.42 | 0.80 | 89 | 0.08 | 0.03 | -55 |
| JE | 1.95 | 1.06 | -46 | 6.21 | 7.21 | 16 |
| SE | 1.81 | 0.97 | -46 | 4.81 | 5.30 | 10 |
| ASM | 0.35 | 0.65 | 88 | 0.02 | 0.01 | -50 |
| IV | 0.04 | 0.02 | -49 | 0.11 | 0.11 | -5 |

**Figure 2:** Representative transaxial baseline neoadjuvant chemotherapy and radiation therapy (NCRT; pre-NCRT) and restaging NCRT (post-NCRT) images and corresponding values of six radiomic features of a pathologic complete responder (Mandard tumor regression grade 1) and a pathologic incomplete responder (Mandard tumor regression grade 5). At diagnosis, the patient who had a pathologic complete response had a T2, N0, M0 adenocarcinoma of the distal esophagus, and the patient who had a pathologic incomplete response had a T3, N0, M0 adenocarcinoma of the distal esophagus. For the patient with a pathologic complete response, joint maximum (JM) and angular second moment (ASM) initially showed high values that increased even more after NCRT, whereas median absolute deviation (MAD), joint entropy (JE), sum entropy (SE), and inverse variance (IV) initially showed low values which decreased even more after NCRT. An inverse trend applied for the pathologic incomplete responder.

these performance measures may potentially be optimistic. To adjust for this optimism, the models were internally validated by bootstrap resampling with 20000 replicates, yielding optimism-corrected Nagelkerke $R^2$ and area under the receiver operating characteristic curve. The constructed prediction model was compared on these terms with six reference prediction models. On the basis of an earlier published article (15), six reference prediction models were constructed composed of pre-NCRT maximum standardized uptake value (model 1); post-NCRT maximum standardized uptake value (model 2); relative difference between pre-NCRT and post-NCRT $^{18}$F-FDG PET maximum standardized uptake value (model 3); clinical T-stage (model 4); clinical T-stage and histologic confirmation (model 5); and clinical T-stage, histologic analysis result, and pre-NCRT long-run low-gray-level emphasis (model 6).

## Results

All patients underwent the full dose of radiation therapy and 72 patients (97.2%) underwent four or five cycles of chemotherapy (Table 1). All volumes of interest had a minimum volume of 10 cm$^3$ as required for textural analysis to provide valuable complementary information (22). The median overall tumor volume before treatment was 52.1 cm$^3$ (interquartile range, 43.8 cm$^3$) and decreased to 41.8 cm$^3$ (interquartile range, 25.6 cm$^3$) at the restaging PET scan. Of all patients, 16 patients (21.9%) obtained a pathologic complete response, whereas 57 patients (78.1%) obtained a pathologic incomplete response.

### Model Construction

Of all studied radiomic features, 86 features were robust for delineation variations, including all traditional radiomic features (exhibiting intraclass correlation coefficients, >0.94). Of these robust radiomic features, 22 pre-NCRT radiomic features, 45 post-NCRT radiomic features, 34 radiomic features that showed relative difference between pre-NCRT and post-NCRT $^{18}$F-FDG PET, and 11 histogram distances were significant at univariable logistic regression analysis (Table E1 [online]). Of the studied clinical parameters, we found a significant association between pathologic complete response and histologic analysis result ($P$ = .01), clinical N-stage ($P$ < .01), and clinical T-stage

**Table 2**

**Estimated Regression Coefficients of Prediction Models 7–12**

| Parameter | Model 7 | | | Model 8 | | | Model 9 | | | Model 10 | | | Model 11 | | | Model 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Regression Coefficient Estimate | SE | *P* Value | Regression Coefficient Estimate | SE | *P* Value | Regression Coefficient Estimate | SE | *P* Value | Regression Coefficient Estimate | SE | *P* Value | Regression Coefficient Estimate | SE | *P* Value | Regression Coefficient Estimate | SE | *P* Value |
| Intercept | −1.28 | 0.28 | <.001 | −1.27 | 0.28 | <.001 | −1.27 | 0.29 | <.001 | 0.85 | 0.84 | .31 | 0.88 | 0.85 | .30 | 0.90 | 0.87 | .30 |
| Clinical T-stage | | | <.01 | | | <.01 | | | <.01 | | | <.01 | | | <.01 | | | <.01 |
| T2 | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | |
| T3 and T4a | −2.81 | 0.95 | | −2.78 | 0.90 | | −2.76 | 0.91 | | −2.70 | 0.92 | | −2.71 | 0.93 | | −2.92 | 0.94 | |
| Joint maximum | 0.83 | 0.38 | .03 | | | | | | | | | | | | | | | |
| Median absolute deviation | | | | −0.74 | 0.68 | .27 | | | | | | | | | | | | |
| Joint entropy | | | | | | | −0.68 | 0.44 | .12 | | | | | | | | | |
| Sum entropy | | | | | | | | | | −0.68 | 0.41 | .09 | | | | | | |
| Angular second moment | | | | | | | | | | | | | 0.74 | 0.37 | .05 | | | |
| Inverse variance | | | | | | | | | | | | | | | | −0.71 | 0.38 | .06 |

Note.—All reported radiomic features were measured at restaging NCRT PET imaging. SE = standard error.

($P$ = .10). Of the traditional radiomic features, overall tumor volume was significant at both pre-NCRT ($P$ = .03) and post-NCRT ($P$ = .03) $^{18}$F-FDG PET, whereas total lesion glycolysis was only significant at post-NCRT $^{18}$F-FDG PET ($P$ = .01). These variables were introduced to the least absolute shrinkage and selection operator regularization process. Least absolute shrinkage and selection operator selected clinical T-stage and post-NCRT joint maximum, which was derived from the gray-level co-occurrence matrix. The joint maximum is the probability corresponding to the most common gray-level co-occurrence in the gray-level co-occurrence matrix. Clinical T-stage and post-NCRT joint maximum were introduced to a logistic regression analysis (model 7). Minimal effects of multicollinearity within the model were found, as quantified by variance inflation factors of 1.01 and 1.01 for the variable clinical T-stage and post-NCRT joint maximum, respectively.

High positive (Spearman $\rho > 0.90$) or negative (Spearman $\rho < -0.90$) correlations were found between post-NCRT joint maximum and five other radiomic features (Fig 1), including the post-NCRT median absolute deviation (Spearman $\rho = -0.90$) and the 4-Gy-level co-occurrence matrix-derived features post-NCRT joint entropy (Spearman $\rho = -0.91$), post-NCRT sum entropy (Spearman $\rho = -0.92$), post-NCRT angular second moment (Spearman $\rho = 0.98$), and post-NCRT inverse variance (Spearman $\rho = -0.91$). We tested whether these radiomic features were redundant in a multivariable regression analysis by exchanging joint maximum by each of these five radiomic features and to refit the respective logistic regression models (models 8–12). In Figure 2, representative pre-NCRT and post-NCRT $^{18}$F-FDG PET images and corresponding values of these six radiomic features are provided for a pathologic complete responder and a pathologic incomplete responder. The regression coefficients of models 1–3 and models 7–12 are shown in Table E2 (online) and Table 2, respectively.

**Table 3**

**Estimates of Model Performance for Prediction Models of Stages M1–11**

| | Goodness-of-Fit | | | Discrimination | | Calibration | | | Validation | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | −2 Log Likelihood | Akaike Information Criterion | Nagelkerke $R^2$ | AUC | Discrimination Slope | Intercept | Slope | Hosmer-Lemeshow *P* Value | Internal Validated Nagelkerke $R^2$ | Internal Validated AUC |
| Model 1 | 76.53 | 80.53 | 0.01 | 0.50 | −0.01 | −0.42 | 0.59 | .05 | 0 | 0.47 |
| Model 2 | 76.66 | 80.66 | 0.00 | 0.34 | −0.01 | −3.40 | −1.59 | | −0.06 | 0.31 |
| Model 3 | 73.83 | 77.83 | 0.06 | 0.43 | 0.03 | 0.93 | 2.08 | .17 | 0.01 | 0.41 |
| Model 4 | 61.52 | 65.52 | 0.29 | 0.70 | 0.20 | −0.53 | 0.79 | | 0.19 | 0.67 |
| Model 5 | 57.83 | 63.83 | 0.35 | 0.75 | 0.25 | −0.19 | 0.86 | >.999 | 0.30 | 0.74 |
| Model 6 | 52.69 | 60.69 | 0.43 | 0.73 | 0.32 | −0.26 | 0.80 | .29 | 0.33 | 0.70 |
| Model 7 | 56.10 | 62.10 | 0.38 | 0.82 | 0.27 | −0.17 | 0.87 | .78 | 0.33 | 0.81 |
| Model 8 | 59.88 | 65.88 | 0.32 | 0.78 | 0.22 | −0.18 | 0.86 | .68 | 0.26 | 0.76 |
| Model 9 | 58.79 | 64.79 | 0.34 | 0.78 | 0.23 | −0.17 | 0.88 | .92 | 0.29 | 0.77 |
| Model 10 | 58.34 | 64.34 | 0.34 | 0.78 | 0.24 | −0.16 | 0.88 | .92 | 0.30 | 0.77 |
| Model 11 | 57.14 | 63.14 | 0.36 | 0.82 | 0.26 | −0.17 | 0.87 | .77 | 0.32 | 0.81 |
| Model 12 | 57.58 | 63.58 | 0.36 | 0.81 | 0.25 | −0.18 | 0.85 | .32 | 0.30 | 0.80 |

Note.—AUC = area under the receiver operating characteristic, model 1 = pre-NCRT $SUV_{max}$, model 2 = post-NCRT $SUV_{max}$, model 3 = relative difference between pre-NCRT and post-NCRT $^{18}$F-FDG PET $SUV_{max}$, model 4 = clinical T-stage, model 5 = clinical T-stage and histologic analysis result, model 6 = clinical T-stage and histologic analysis result and pre-NCRT long-run low-gray-level emphasis, model 7 = clinical T-stage and post-NCRT joint maximum, model 8 = clinical T-stage and post-NCRT median absolute deviation, model 9 = clinical T-stage and post-NCRT joint entropy, model 10 = clinical T-stage and post-NCRT sum entropy, model 11 = clinical T-stage and post-NCRT angular second moment, model 12 = clinical T-stage and post-NCRT inverse variance, post-NCRT = restaging NCRT, pre-NCRT = baseline NCRT, $SUV_{max}$ = maximum standardized uptake value.

### Model Performance

The performance measures of models 7–12 are shown in Table 3. Model 7 exhibited the following goodness-of-fit metrics: a −2 log likelihood of 56.10, an Akaike information criterion of 62.10, and a Nagelkerke $R^2$ of 0.32. The model had a good discriminatory accuracy, with an area under the receiver operating characteristic curve of 0.82 and a discrimination slope of 0.27. The model was well calibrated, with an intercept of −0.17, a slope of 0.87, and a Hosmer-Lemeshow *P* value of .78 (Fig 3). After internal validation, the Nagelkerke $R^2$ and area under the receiver operating characteristic curve slightly decreased to 0.33 and 0.81, respectively. Models 1–5 all exhibited worse goodness-of-fit, discrimination, and calibration compared with model 7. Model 6 had a better goodness-of-fit than model 7, however, this was not reflected by a higher discriminatory accuracy and a better calibration. The model performances of 7–12 were relatively consistent, though model 7 did exhibit the best goodness-of-fit and discrimination, which also persisted after internal validation. Although models 9–10 had a slightly better calibration than model 7, the calibrations of models 7–12 were also relatively consistent.

## Discussion

Clinical evaluation of response to NCRT is important for patients who have potentially curable locally advanced esophageal cancer for exploring future personalized treatment and for prediction of prognostic outcome. Recently, response increasingly was defined on the basis of tumor biology rather than on tumor morphologic analysis by using Response Evaluation Criteria in Solid Tumors (23,24). Image quantification with radiomics is an emerging noninvasive approach to predict survival outcome and response in cancer treatment. In our study, we constructed a prediction model on the basis of clinical evaluation of tumor invasion depth (T-stage) and post-NCRT joint maximum. Joint maximum is a measure of orderliness, quantifying the systematic arrangement of voxel intensity differences. Joint entropy and angular second moment are other radiomic features that are measures of image orderliness. High values of joint maximum and angular second moment occur in orderly images, whereas large joint entropy values indicate disorder. For each increase in joint maximum, the odds of pathologic complete response increases by a factor of $e^{\text{Coef}_{\text{joint maximum}}} = e^{0.83} = 2.3$, where $e$ is the Euler number and $Coef_{jointmaximum}$ is the regression coefficient estimate of the joint maximum. In accordance with our expectations, this would mean that higher post-NCRT $^{18}$F-FDG PET orderliness increases the probability for achieving pathologic complete response.

The final prediction model was superior compared with the reference prediction models composed of conventional maximum standardized uptake value measurements (models 1–3) and clinical parameters (models 4–5). Moreover, it demonstrated a higher discriminatory accuracy than did the reference model, consisting of clinical parameters along with a pre-NCRT radiomic feature (models 6), which suggested the higher potential of post-NCRT radiomic features in predicting pathologic complete response in patients with esophageal

**Figure 3**

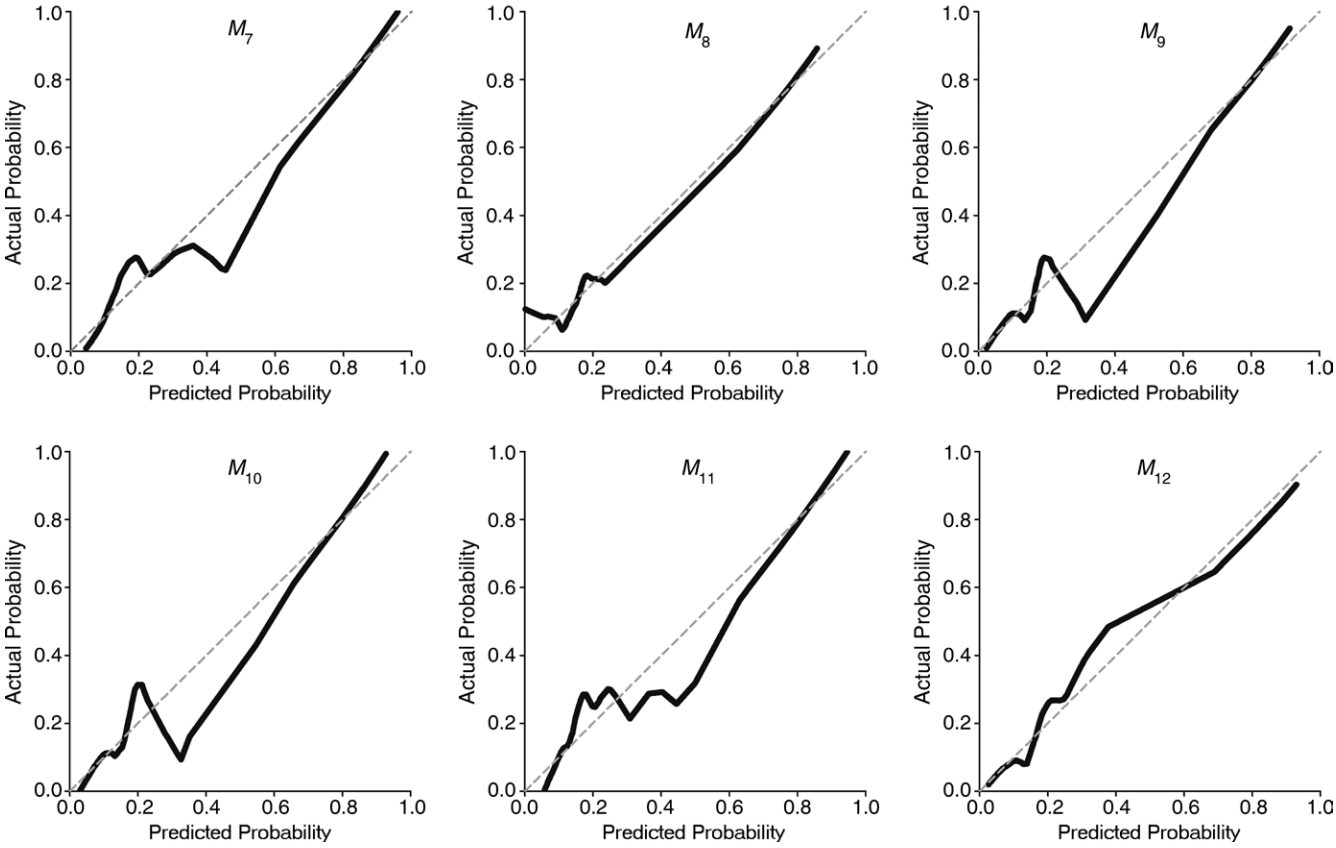



**Figure 3:** Calibration plots of models 7–12 referring to the agreement between the predicted probability of pathologic complete response by models 7–12 and the true (observed) probability of pathologic complete response (solid line). The dashed line indicates perfect calibration.

cancer. This higher discriminatory accuracy was retained when post-NCRT joint maximum was exchanged by one of the correlated radiomic features. Predictions on the basis of relative differences of radiomic features and histogram distances were clinically not relevant.

Over the years, numerous radiomic features were reported as promising. To validate whether these radiomic features also provide complementary clinical value in external cohorts, pooling those results is required. In a systematic Medline database search, we found 10 relevant studies that investigated the value of PET/CT radiomic features in the prediction of therapy response in esophageal cancer (11–18,25,26). The study by Van Rossum et al (11) was, to our knowledge, the only study that performed quantitative $^{18}$F-FDG PET from pre-NCRT and post-NCRT $^{18}$F-FDG PET, classified response as complete (absence of viable tumor tissue) versus incomplete (any grade of residual tumor tissue), and performed a univariable logistic regression analysis. Consistent with our results, Van Rossum et al reported that the post-NCRT radiomic features joint maximum, joint entropy, sum entropy, and angular second moment were significant at univariable logistic regression analysis. However, inverse variance was reported as nonsignificant and median absolute deviation was not analyzed. They proposed four multivariable prediction models, of which the prediction model exhibiting the highest discriminatory accuracy consisted of 10 different clinical and (subjective and quantitative) $^{18}$F-FDG PET parameters. Their model contained substantially more variables and hence was more complex, and their model exhibited a slightly lower optimism-corrected area under the receiver operating characteristic curve of 0.77 compared with 0.81 of our proposed prediction model.

Though promising, the use of radiomics to select patients for different treatment strategies on the basis of identification of response is still not practicable because of numerous unsolved problems and the complexity of data. There is a significant lack of standardization including the use of different scanners and image acquisition protocols among hospitals and different feature extraction workflows, which could lead to a large variability in radiomic features (27). Moreover, not all textural features are deemed stable enough with respect to the effect of different devices and delineation variation. The study by Hatt et al (28) showed that robust radiomic features for delineation variation included joint entropy, inverse difference,

dissimilarity, and zone percentage; however, they assessed a limited number of radiomic features. Therefore, our study tested the effect of delineation variation on all studied radiomic features by creating multiple artificial tumor delineations. Our results were generally consistent with the study by Hatt et al. Post-NCRT tumor delineation is complicated in complete responders because of the absence of viable tumor tissue. In those patients, the baseline tumor delineation was registered with the restaging PET and the original tumor dimensions were retained. Moreover, localization of the tumor at restaging PET can be difficult because the metabolically active area is often contaminated with NCRT-induced esophagitis.

A shortcoming of our study was the lack of a patient cohort for external validation. However, verification of the proposed prediction models requires large cohorts of homogeneously staged and treated patients. Another key factor that hampers clinical implementation of the proposed prediction model is that there is limited information regarding the repeatability of the proposed radiomic features. In one of the few studies regarding this subject, Tixier et al (29) compared textural features of double-baseline $^{18}$F-FDG PET scans and found that joint entropy exhibited a high reproducibility, whereas angular second moment was characterized by lower reproducibility. Joint maximum, median absolute deviation, sum entropy, and inverse variance were not evaluated.

In summary, higher post-NCRT $^{18}$F-FDG PET orderliness measured by joint maximum increased the probability for achieving pathologic complete response after NCRT in patients with esophageal cancer. In this study, a prediction model composed of clinical T-stage and post-NCRT joint maximum seemed to add important information to the visual PET/CT evaluation to guide eventual omission of surgery for patients who achieve a pathologic complete response.

**Disclosures of Conflicts of Interest: R.J.B.** disclosed no relevant relationships. **J.B.H.** disclosed no relevant relationships. **V.E.M.M.** disclosed no relevant relationships. **W.N.** disclosed no relevant relationships. **G.K.U.** disclosed no relevant relationships. **R.H.J.A.S.** disclosed no relevant relationships. **J.T.M.P.** disclosed no relevant relationships.

## References

1. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med 2012;366(22):2074–2084.
2. Shapiro J, van Lanschot JJB, Hulshof MCCM, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol 2015;16(9):1090–1098.
3. van Hagen P, Wijnhoven BP, Nafteux P, et al. Recurrence pattern in patients with a pathologically complete response after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. Br J Surg 2013;100(2):267–273.
4. Meguid RA, Hooker CM, Taylor JT, et al. Recurrence after neoadjuvant chemoradiation and surgery for esophageal cancer: does the pattern of recurrence differ for patients with complete response and those with partial or no response? J Thorac Cardiovasc Surg 2009;138(6):1309–1317.
5. Smit JK, Güler S, Beukema JC, et al. Different recurrence pattern after neoadjuvant chemoradiotherapy compared to surgery alone in esophageal cancer patients. Ann Surg Oncol 2013;20(12):4008–4015.
6. Zanoni A, Verlato G, Giacopuzzi S, et al. Neoadjuvant concurrent chemoradiotherapy for locally advanced esophageal cancer in a single high-volume center. Ann Surg Oncol 2013;20(6):1993–1999.
7. Lordick F, Ott K, Krause BJ, et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. Lancet Oncol 2007;8(9):797–805.
8. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of $^{18}$F FDG PET: a systematic review. Radiology 2010;254(3):707–717.
9. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. J Nucl Med 2012;53(7):1041–1047.
10. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res 2015;21(2):249–257.
11. van Rossum PS, Fried DV, Zhang L, et al. The Incremental Value of Subjective and Quantitative Assessment of $^{18}$F-FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer. J Nucl Med 2016;57(5):691–700.
12. 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(5):1375–1382.
13. Tan S, Zhang H, Zhang Y, Chen W, D'Souza WD, Lu W. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in $^{18}$F-FDG uptake patterns. Med Phys 2013;40(10):101707.
14. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline $^{18}$F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52(3):369–378.
15. Beukinga RJ, Hulshoff JB, van Dijk LV, et al. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment 18F-FDG PET/CT imaging. J Nucl Med 2017;58(5):723–729.
16. Yip SS, Coroller TP, Sanford NN, Mamon H, Aerts HJ, Berbeco RI. Relationship between the Temporal Changes in Positron-Emission-Tomography-Imaging-Based Textural Features and Pathologic Response and Survival in Esophageal Cancer Patients. Front Oncol 2016;6:72.
17. Nakajo M, Jinguji M, Nakabeppu Y, et al. Texture analysis of $^{18}$F-FDG PET/CT to predict tumour response and prognosis of patients with esophageal cancer treated by chemoradiotherapy. Eur J Nucl Med Mol Imaging 2017;44(2):206–214.
18. Zhang H, Tan S, Chen W, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal $^{18}$F-FDG PET features, clinical parameters, and demographics. Int J Radiat Oncol Biol Phys 2014;88(1):195–203.
19. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. Cancer 1994;73(11):2680–2686.
20. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guide-

lines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging 2015;42(2):328–354.

21. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - feature definitions. CoRR 2016; abs/1612.07003. https://arxiv.org/abs/1612.07003v5. Accessed February 1, 2018.

22. Hatt M, Majdoub M, Vallières 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(1):38–44.

23. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45(2):228–247.

24. Oxnard GR, Schwartz LH. Response phenotype as a predictive biomarker to guide treatment with targeted therapies. J Clin Oncol 2013;31(30):3739–3741.

25. Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. PLoS One 2017;12(3):e0173208.

26. Ypsilantis PP, Siddique M, Sohn HM, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PLoS One 2015;10(9):e0137036.

27. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49(7):1012–1016.

28. 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(11):1662–1671.

29. 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(5):693–700.