Prediction of Non-Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer Patients with 18F-FDG PET Radiomics Based Machine Learning Classification

Beukinga, Roelof J; Poelmann, Floris B; Kats-Ugurlu, Gursah; Viddeleer, Alain R; Boellaard, Ronald; De Haas, Robbert J; Plukker, John Th M; Hulshoff, Jan Binne

Published in:
Diagnostics

DOI:
10.3390/diagnostics12051070

diagnostics

MDPI

Article

Prediction of Non-Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer Patients with ^{18}F-FDG PET Radiomics Based Machine Learning Classification

Roelof J. Beukinga [1,†], Floris B. Poelmann [2,†], Gursah Kats-Ugurlu [3], Alain R. Viddeleer [4], Ronald Boellaard [1,5], Robbert J. de Haas [4], John Th. M. Plukker [2] and Jan Binne Hulshoff [1,4,*]

1 Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB Groningen, The Netherlands; j.beukinga@gmail.com (R.J.B.); r.boellaard@umcg.nl (R.B.)
2 Department of Surgical Oncology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB Groningen, The Netherlands; florispoelmann@gmail.com (F.B.P.); j.t.m.plukker@umcg.nl (J.T.M.P.)
3 Department of Pathology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB Groningen, The Netherlands; g.kats-ugurlu@umcg.nl
4 Department of Radiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9700 RB Groningen, The Netherlands; a.r.viddeleer@umcg.nl (A.R.V.); r.j.de.haas@umcg.nl (R.J.d.H.)
5 Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands
* Correspondence: jb.hulshoff@umcg.nl
† These authors contributed equally to this work.

check for updates

Citation: Beukinga, R.J.; Poelmann, F.B.; Kats-Ugurlu, G.; Viddeleer, A.R.; Boellaard, R.; de Haas, R.J.; Plukker, J.T.M.; Hulshoff, J.B. Prediction of Non-Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer Patients with ^{18}F-FDG PET Radiomics Based Machine Learning Classification. *Diagnostics* **2022**, *12*, 1070. https://doi.org/10.3390/diagnostics12051070

Academic Editor: Manuel Scimeca

Received: 17 March 2022
Accepted: 22 April 2022
Published: 24 April 2022

Abstract: Background: Approximately 26% of esophageal cancer (EC) patients do not respond to neoadjuvant chemoradiotherapy (nCRT), emphasizing the need for pre-treatment selection. The aim of this study was to predict non-response using a radiomic model on baseline ^{18}F-FDG PET. Methods: Retrospectively, 143 ^{18}F-FDG PET radiomic features were extracted from 199 EC patients (T1N1-3M0/T2–4aN0-3M0) treated between 2009 and 2019. Non-response (n = 57; 29%) was defined as Mandard Tumor Regression Grade 4–5 (n = 44; 22%) or interval progression (n = 13; 7%). Randomly, 139 patients (70%) were allocated to explore all combinations of 24 feature selection strategies and 6 classification methods towards the cross-validated average precision (AP). The predictive value of the best-performing model, i.e AP and area under the ROC curve analysis (AUC), was evaluated on an independent test subset of 60 patients (30%). Results: The best performing model had an AP (mean ± SD) of 0.47 ± 0.06 on the training subset, achieved by a support vector machine classifier trained on five principal components of relevant clinical and radiomic features. The model was externally validated with an AP of 0.66 and an AUC of 0.67. Conclusion: In the present study, the best-performing model on pre-treatment ^{18}F-FDG PET radiomics and clinical features had a small clinical benefit to identify non-responders to nCRT in EC.

Keywords: esophageal neoplasms; neoadjuvant therapy; positron-emission tomography

1. Introduction

Most patients with locally advanced esophageal cancer (T1N1-3M0/T2–4aN0-3M0) benefit from neoadjuvant chemoradiotherapy (nCRT) followed by esophagectomy [1]. After nCRT, 29% of these patients have a pathologically complete response and 32% have a near-complete response with < 10% vital tumor cells [1]. However, a substantial group of patients does not respond: 8% develop progressive disease usually as interval metastases, and 18% only achieve a limited response (i.e., >50% remaining vital tumor cells) [1,2]. Patients with poor response to nCRT followed by complete resection have a similar prognosis as those who undergo primary esophagectomy [3]. Pre-treatment identification

of non-responders would allow for alternative treatment strategies, e.g., earlier surgical intervention or additional targeted therapies. This custom-based approach may prevent prolonged useless exposure to nCRT with potential risk of radiation-induced toxicity or tumor expansion with delay of surgery.

There is increasing evidence that intratumoral heterogeneity is a major determinant of non-response to nCRT. Heterogeneity on the cellular level can be caused by genetically distinct subpopulations for sustained tumor growth, including cancer stemness, genetic diversity in ligand/receptor expression, tumor microenvironment with metabolic reprogramming, and epigenetic alterations [4–6]. As such, it may be caused by distinct subclonal populations with specific patterns of oxygen consumption, glucose metabolism, and cellular proliferation as reflected by subtle spatial variations on medical images [6–8]. Although these variations are difficult to detect during regular radiological reading, they may be revealed by voxel-wise pattern recognition techniques such as radiomics [6–9]. Radiomics phenotyping is a non-invasive analysis which encompasses image acquisition followed by high-throughput extraction of quantitative features from regions-of-interest defined on medical images. These predefined features capture geometric, intensity, and textural information about the tumor and provide a huge amount of non-invasive imaging biomarkers which can be modeled using machine learning classifiers to predict treatment response and prognosis.

Several studies reported promising results in predicting complete response to nCRT when using pre- and/or post-nCRT radiomics derived from CT and/or ^{18}F-FDG PET [10–19]. The innovative field of radiomics might provide similar opportunities in managing non-responding esophageal cancer patients. A great benefit using ^{18}F-FDG PET radiomics is the unique ability to provide whole-body quantitative information of spatial phenotypic variation in metabolism, thereby capturing tumor site-related information about tumor resistance to nCRT such as hypoxia, necrosis, and cellular proliferation [20].

However, studies on radiomics are based on complex statistics and analyses, and truly clinically relevant findings are still lacking. Since it is important to transfer knowledge from scientific research more early into real time practice, defining current position using a relative large number of uniform staged EC patients could add to a critical sound view before being suitable in prospective studies. Therefore, the aim of this study was to construct a useful model combining clinical information and radiomic features from pretreatment ^{18}F-FDG PET scans to predict non-response to nCRT in esophageal cancer.

2. Materials and Methods

2.1. Patients

This retrospective study was conducted in accordance with the Dutch guidelines for retrospective studies and rules of the local institutional ethical board, the local ethical board waived the requirement to obtain informed consent (METc 202000093). Between January 2009 and August 2019, 199 patients with locally advanced esophageal cancer (T1N1-3M0/T2–4aN0-3M0) treated with nCRT at our institution were included (Figure 1 displays the inclusion and exclusion criteria). Data collection and reporting of analysis was performed according to the STARD guidelines.

2.2. Staging and Treatment

Patients were staged with a thoraco-abdominal CT (Biograph mCT 4–64 PET/CT; Siemens, Erlangen, Germany), ^{18}F-FDG PET/CT (Biograph mCT-64 PET/CT; Siemens, Knoxville, TN, USA), and endoscopic ultrasound. Patients were discussed in the multidisciplinary upper gastrointestinal tumor board and treated according to the CROSS regimen (5 cycles of carboplatin (2 mg·min·mL^{-1}) and paclitaxel (50 mg/m^2) with 41.4 Gy in 23 fractions) [1]. Restaging was performed 6–8 weeks after nCRT with CT (before 2014) or ^{18}F-FDG PET/CT (after 2014). Surgical treatment consisted of a minimally invasive or open transthoracic esophagectomy.

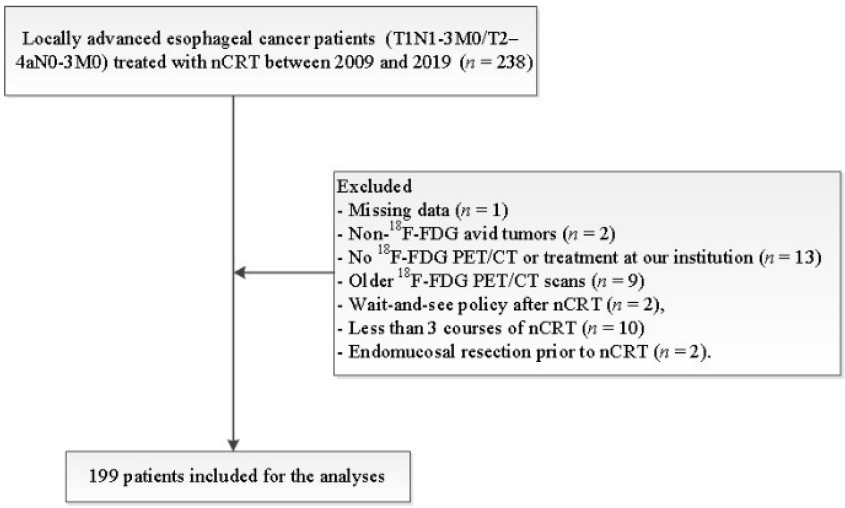

Figure 1. Inclusion and exclusion flowchart. Abbreviations: nCRT = neoadjuvant chemoradiotherapy.

2.3. Histopathologic Response Evaluation

Two experienced pathologists assessed the resected surgical specimen and scored response on the five-point Mandard tumor regression grade (TRG) scale. Patients were considered non-responders if residual tumor was scored as TRG 4 (i.e., fibrosis and tumor cells with preponderance of tumor cells), as TRG 5 (i.e., tumor tissue without signs of regression), or if progressive disease was detected at restaging or during surgery [21].

2.4. PET/CT Imaging

After at least six hours of fasting, all patients received 3 MBq/kg ^{18}F-FDG 60 min prior to imaging. Low-dose CT (80–120 kV; 20–35 mAs; and 5 mm section thickness) and PET images (voxel size 3.1819 mm × 3.1819 mm × 2 mm and 2–3 min scans per bed position) were acquired in radiation treatment planning position. To harmonize SUV (standardized uptake value), images were reconstructed in compliance with either NEDPAS or EARL protocols [22].

2.5. Tumor Delineation and Radiomic Feature Extraction

Primary tumors were initially delineated on axial images of the baseline PET scans using SUV thresholding with an in-house delineation tool built in MeVisLab (MeVis Medical Solutions AG, Bremen, Germany; version 3.1.1). To optimize the quality of the delineations, the volume of interest (VOI) was manually corrected using both the CT and PET scan in consensus between the collaborating investigators (RJB and JBH). PET scans and corresponding VOI were resampled to isotropic voxel-dimensions of 2 mm × 2 mm × 2 mm using trilinear interpolation. The interpolated VOIs were rounded to binary images. From each VOI, 143 ^{18}F-FDG PET-derived radiomic features were extracted with software developed in Matlab 2018b (MathWorks Inc, Natick, MA, USA). Image processing and feature extraction were performed in compliance with guidelines provided by the Image Biomarker Standardization Initiative [23]. To reduce both computational workload and the impact of image noise, textural features were extracted from discretized image stacks ($X\_discretized = \lfloor X\_SUV/0.25 \rfloor + 1$). From these discretized image stacks, the spatial distribution of gray-level intensities was scored in three dimensions (26-voxel connectivity) into a single merged texture matrix from which the textural features were calculated.

2.6. Radiomics Machine Learning Pipeline

Figure 2 shows the machine learning pipeline written in Python 3 using the open-source machine learning library Scikit Learn (version 0.22.1). Radiomic features entered the machine learning pipeline together with clinical features (histology, clinical T- and N-stage). All continuously scaled features were normalized. If a feature distribution had

a skewness between −0.5 and 0.5, the feature was robustly normalized by removing the median and scaling to the interquartile range. In case of an absolute skewness > 0.5, the feature was transformed by a Yeo-Johnson power transformation (which applies monotonic transformations to make the data more Gaussian-like). The data were randomly divided into a training (70% of the samples) and validation test subset (30% of the samples), with preservation of the original response distribution.

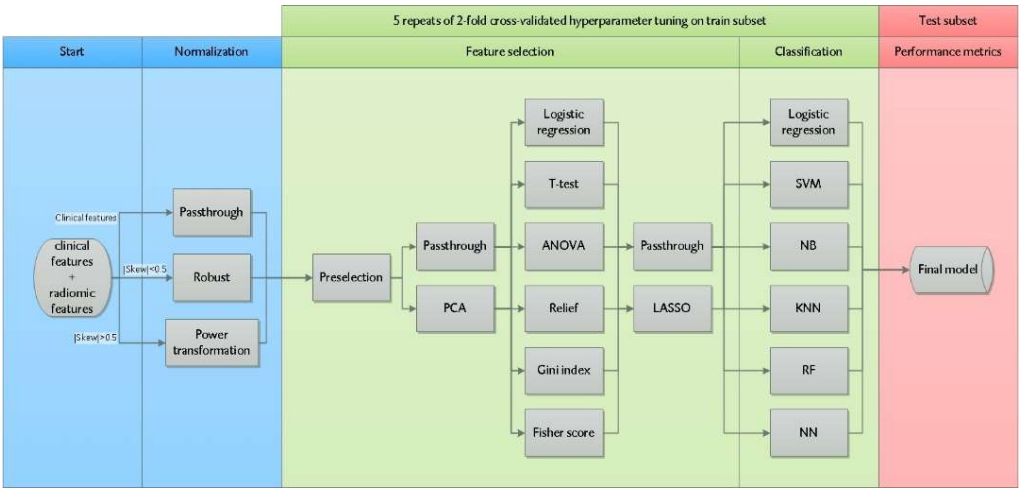

Figure 2. Radiomics machine learning pipeline to train and select a model predicting non-response to nCRT. Radiomic and clinical features were normalized up front (blue area). Hyperparameter tuning was performed on the training subset (green area) with 24 unique feature selection strategies and 6 classification methods. The model with the highest mean average precision (AP) over the different cross validation folds was selected. The performance of this model was tested on the test subset (red area). Abbreviations: Skew = skewness of the distribution, SVM = support vector machine, NB = Gaussian Naive Bayes, KNN = K-nearest neighbors, RF = random forest, and NN = neural network.

To prevent overfitting, i.e., when the model unintentionally is learning noise instead of the underlying trend of the data, the feature space was reduced in four consecutive feature selection steps as illustrated in Figure 2 (resulting in 24 unique feature selection strategies). All combinations of these feature selection strategies were explored with six classification methods (resulting in 144 different machine learning strategies) in terms of model performance on the training and test subsets. In the first feature selection step, we eliminated radiomic features with a low multivendor reproducibility identified by earlier research (i.e., intraclass correlation coefficient ICC < 0.6) and radiomic features with a high Pearson correlation ($\rho > 0.8$) with conventional features (volume, SUVmax, SUVpeak, SUVmean, and/or total lesion glycolysis) [24]. In the second feature selection step, principal component analysis was used to further reduce dimensionality by creating new uncorrelated variables (i.e., principal components) from the original dataset. The first principal components (sorted by the amount of variance in the data), which explained > 95% of the total variance in the data, were selected for further analysis. Thirdly, to rank features, the impact of 6 univariable filter methods (logistic regression, ANOVA, Fisher score, Relief, T-score, and Gini index) was investigated. The final feature selection step involved incorporation of the least absolute shrinkage and selection operator (LASSO), a regularization technique which simultaneously prevents overfitting and performs feature selection. After feature selection, six different machine learning classifiers were trained: logistic regression, support vector machine, random forest, Gaussian naive Bayes, neural network, and K-nearest neighbors. A 2-fold cross-validation was repeated 5 times in the training subset to tune hyperparameters of the filter methods, regularization, and classifiers, and to select the best-performing model based on the mean average precision (AP) over the different folds. AP measures the area under the precision–recall curve, which describes the trade-off

between precision (i.e., positive predictive value) and recall (i.e., sensitivity). The AP metric is particularly useful in this study as it only considers the positive class (minority of the cases) and is unconcerned of the true negatives (majority of the cases). The AP adds statistical value when it exceeds the percentage of non-responding patients in the test subset. To further improve the performance, we used a soft-voting rule classifier which aggregates the predictions of the 10 best-performing models by averaging the class–probabilities of these models. The generalization performances of all models were evaluated on the independent test subset.

3. Results

3.1. Patients Characteristics

Patient and tumor characteristics of the response and non-response group are summarized in Table 1. Among the included 199 patients, 57 (29%) were non-responders; 39 (68%) had TRG 4 and 5 (9%) TRG 5. Progressive disease was detected at restaging in ten patients (5%) and intraoperatively in three (2%) patients. All these 13 patients were not amenable to further surgery and were considered as having $\geq$ TRG 4 based on macroscopic progressive tumor. Of all patients, 139 (70%) were allocated to the training and 60 (30%) to the test subset. There were no significant differences in patient characteristics between the training and test subset.

Table 1. Patient and tumor characteristics of responders versus non-responders.

Characteristic	Response (n = 142) n (%)	Non-Response (n = 57) n (%)	p-Value [1]
Gender (Male)	113 (79.6)	48 (84.2)	0.446
Age (years), median (IQR)	66 (61–71)	67 (61–72)	0.546 [2]
Histology Adenocarcinoma Squamous cell carcinoma	 124 (87.3) 18 (12.7)	 53 (93.0) 4 (7.0)	0.231
Tumor location Mid Distal Gastroesophageal junction	 20 (14.1) 96 (67.6) 26 (18.3)	 2 (3.5) 42 (73.7) 13 (22.8)	0.057
Tumor length (cm), median (IQR)	6.0 (4.0–7.0)	5.0 (4.0–8.0)	0.595 [2]
Clinical T-stage T1 T2 T3 T4a	 2 (1.4) 28 (19.7) 107 (75.4) 5 (3.5)	 0 (0.0) 8 (14.0) 44 (77.2) 5 (8.8)	0.246
Clinical N-stage N0 N1 N2 N3	 30 (21.1) 75 (52.8) 33 (23.2) 4 (2.8)	 16 (28.1) 23 (40.4) 15 (26.3) 3 (5.3)	0.399
CRM (0 mm) R1 NA [3]	 5 (3.5) 0 (0.0)	 3 (5.3) 13 (22.8)	0.371

Abbreviations: IQR = interquartile range, CRM = circumferential resection margin, R0 = microscopically tumor-free resection, R1 = microscopically irradical resection, and NA = not applicable. [1] Likelihood ratio test. [2] Mann–Whitney U test. [3] No resection was performed due to distant metastases found before or during surgery.

3.2. Feature Normalization and Preselection

In total, 143 radiomic and 3 clinical features (clinical T- and N-stage, and tumor histology) entered the machine learning pipeline. In the preselection step, 22 of the 143 extracted radiomic features had a low multivendor reproducibility according to the definitions used in previous research (ICC < 0.6) and were eliminated [24]. Among the remaining 121 radiomic features, 25 had an approximately symmetric distribution (skewness between −0.5 and 0.5) and 96 had a moderate to high skew distribution (absolute skewness > 0.5) and were normalized according to the predetermined normalization approach. After normalization, 65 radiomic features were considered redundant and subsequently removed because of a high Pearson correlation ($\rho > 0.8$) with one or more conventional features (volume, SUVmax, SUVpeak, SUVmean, and total lesion glycolysis), leaving 3 clinical and 56 radiomic features for further analysis. Supplementary Table S1 displays the excluded vendor-dependent and -redundant features. The degree of redundancy between the remaining 56 radiomic features is demonstrated by a correlation heatmap with dendrograms (Figure 3). Of these features, 84% had at least one absolute pair-wise Pearson correlation > 0.8, indicating a substantial amount of feature redundancy remaining in the dataset. This redundancy was attempted to be reduced by the subsequent feature selection steps in the machine learning pipeline.

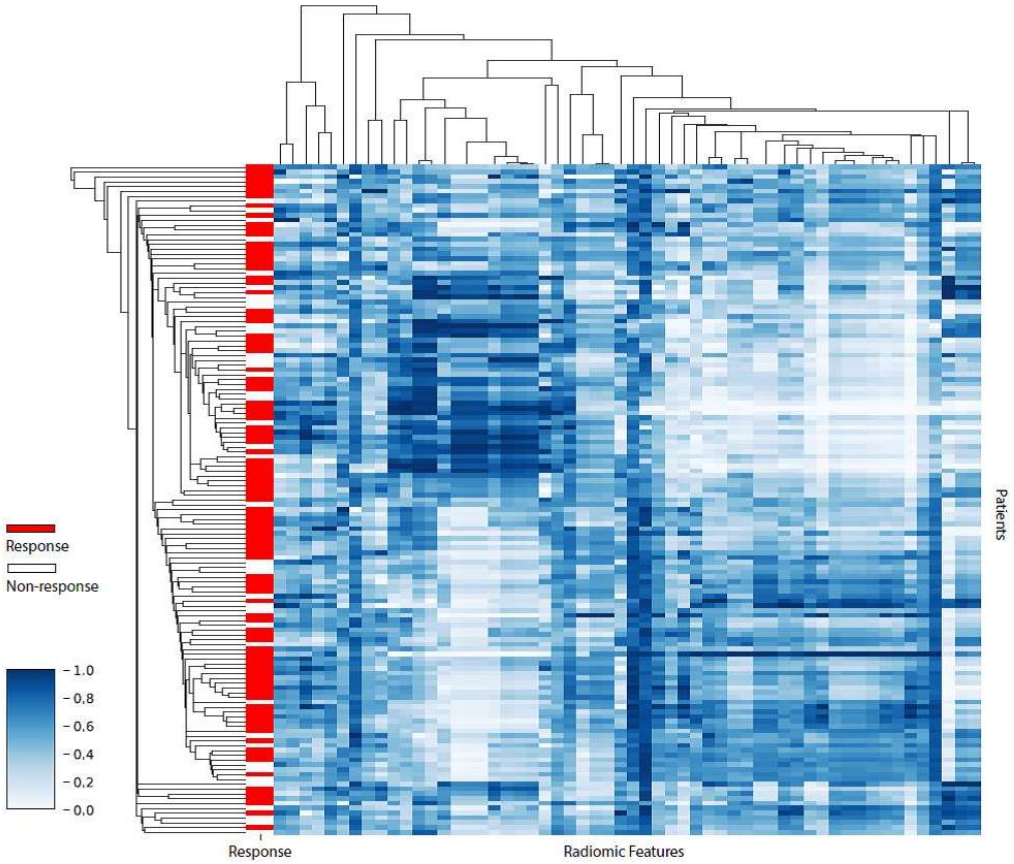

Figure 3. Heatmap revealing radiomic feature clusters with similar expression (standardized on white-blue gradient scale) using unsupervised clustering with Pearson correlation as a measure of similarity. The x-axis represents the preselected radiomic features (n = 56) and the y-axis represents esophageal cancer patients in the training subset (n = 139). The heatmap reveals a substantial amount of feature redundancy.

3.3. Model Selection and Performance

Figure 4 shows the cross-validated model performance of the 10 best performing models selected from the training subset, based on the AP metric. Model 1–4 were essentially identical models with the same features and identical performances but were constructed

through four different machine learning strategies. All these models were support vector machine classifiers trained on the same five principal components, generated by principal component analysis during feature selection. However, these principal components were selected by four different feature selection methods, i.e., relief, ANOVA, logistic regression, and T-score. These models showed an AP (mean ± SD) of 0.47 ± 0.06 on the training subset and were externally validated in the test subset with an AP of 0.66 (Figure 5) and an area under the ROC curve (AUC) of 0.67. This AP is substantially higher than the AP of random classification (percentage non-responding patients in the test subset = 0.28). The learning curve in Figure 6 shows that the training and test AP scores did not fully converge to a point of stability yet, and therefore the model would slightly benefit from more training data. The soft-voting rule classifier, aggregating the predictions from the 10 best-performing models, showed an AP of 0.64 and an AUC of 0.68 on the test subset.

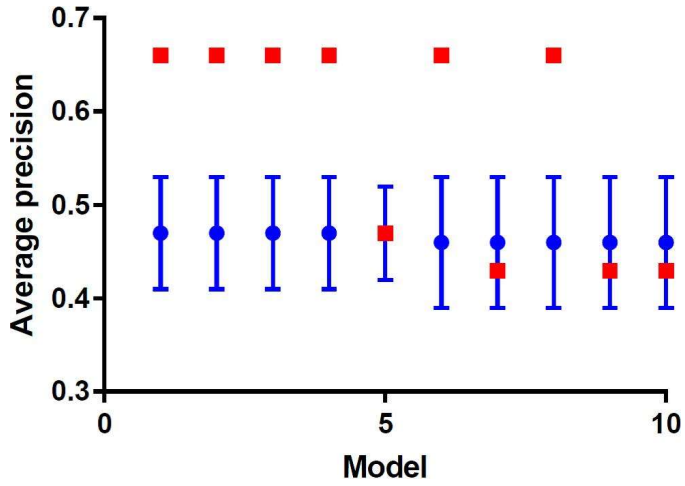

Figure 4. Plot of the 10 best-performing models ordered by the mean average precision over the validation runs in the training subset (blue). The test performance was evaluated on an independent test set (red).

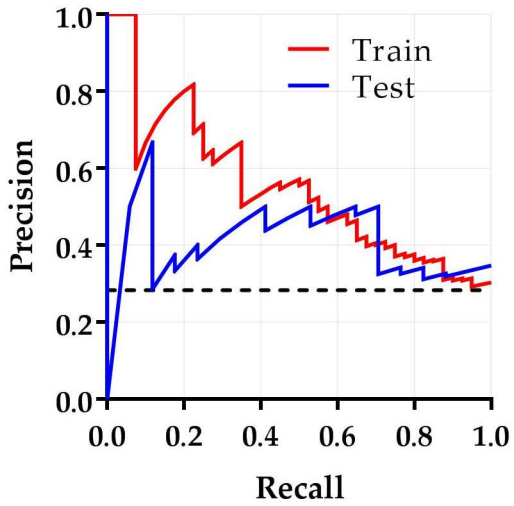

Figure 5. Precision–recall curve of the best performing model demonstrating the trade-off between precision and recall. The area under the precision–recall curve is reflected by the average precision. The average precisions for the training and test subset are 0.47 and 0.66, respectively. The black dashed line is the score of a random classification (0.28).

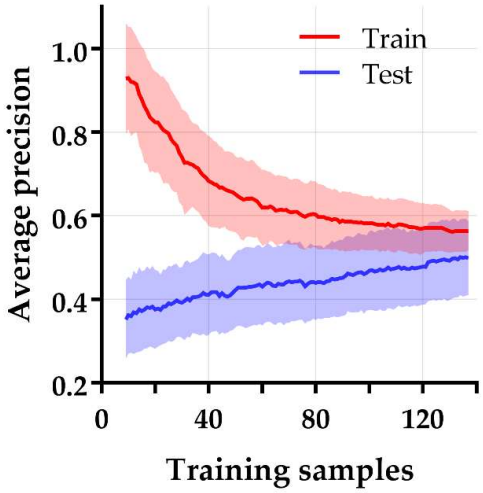

Figure 6. Learning curve of the best-performing model for prediction of non-response after nCRT in esophageal cancer. The average precision is plotted on the y-axis and the number of training samples on the x-axis. The training and test average precision scores did not fully converge to a point of stability yet, suggesting that the training process may slightly benefit from a larger sample size.

To determine the effect of data heterogeneity, additional sub-analyses were performed on clinical T-stage and histology. Patients with clinical stage T1–T2 tumors (n = 38) were excluded because radiomic features from smaller volumes are known to be less reliable [25]. The best performing model in the T3–4a patient group showed a train-AP of 0.47 ± 0.10, a test-AP of 0.48, and a test-AUC of 0.67. Moreover, to increase homogeneity, a separate sub-analysis with only adenocarcinoma patients (n = 177) was executed. In this group, the best-performing model exhibited a train-AP of 0.58 ± 0.09, a test-AP of 0.29, and a test-AUC of 0.46.

4. Discussion

Following promising results of ^{18}F-FDG PET radiomics studies on the prediction of pathologically complete response in esophageal cancer, adequate discriminative ability would be expected in predicting non-response [13,14]. After investigating a wide spectrum of machine learning techniques including data dimension reduction techniques, classifiers, and cross-validated model training, our best performing prediction model was able to learn representative patterns in the dataset (AP 0.66). To test the clinical relevance of this model, only high precisions should be considered within the current clinical scenario as it is extremely important to prevent refrainment of effective nCRT due to false positive predictions. However, the trade-off between recall and precision in this study shows that it is not possible to increase precision without substantially reducing the recall (Figure 6). This would implicate an increase in the number of responding patients that are falsely classified as non-responding patients. Although this study shows a relatively small clinical benefit of combining clinical and radiomic features from pretreatment ^{18}F-FDG PET scans, the predictive power is too low to be clinically applicable in predicting non-response to nCRT in esophageal cancer.

We attempted to find the underlying cause of the relatively low predictive ability of this model. The learning curve in Figure 6 indicates that the training process was halted rather prematurely and may slightly benefit from a larger sample size. Furthermore, despite this study was conducted on a relatively homogeneous patient group, two separate analyses were performed to rule out potential influence of remaining data heterogeneity. First, we limited analyses to T3-T4a tumors to reduce the effect of partial volume effects and possible delineation inconsistencies in smaller cancers. Despite this approach being consistent with earlier studies stating that the complementary information of radiomic features substantially increases with larger volumes [26], it did not improve the model

performance. Moreover, a separate analysis was performed on adenocarcinomas alone, which respond poorer to nCRT than squamous cell carcinomas [1]. However, this subgroup analysis did not reveal any model improvement either.

So far, several studies investigated temporal changes in ^{18}F-FDG PET radiomics features [16,17]. However, as this information can only be extracted after definitive treatment, it has little clinical impact on changing patient management. Tixier et al. did report differences in baseline ^{18}F-FDG PET radiomics between non-responders and partial responders, but this study had a small sample size with no external validation [15]. As already known, the difference between training and test performance results emphasizes the necessity of an external validation group and sufficient sample size in order to determine the true predictive value of radiomic models. In addition, differences in imaging features between the groups might be related to only a small but substantial part of distinct subclones, reflecting a crucial area of tumor biology with genomics driven differences.

One of the main issues of radiomics are unestablished measurement errors (i.e., repeatability, reliability, and reproducibility). Moreover, the majority of ^{18}F-FDG PET radiomics are sensitive to different sources of variation such as the delineation method, image acquisition, or reconstruction protocols [27–29]. In accordance with prior research, a wide range of radiomic features were harmonized by acquiring all scans in a single-center and by using single-vendor settings according to either the "European Association of Nuclear Medicine Research Ltd." (EARL) compliant reconstruction protocols or "Netherlands protocol for standardization of ^{18}F-FDG whole-body PET studies in multi-center trials" (NEDPAS) [22,24]. Due to the retrospective nature of this study, the reconstruction protocol was updated during the course of the study. Only ^{18}F-FDG PET radiomic features reliable in a multi-center and multi-vendor setting were preselected for further analysis. Moreover, radiomic features are sensitive to inconsistent tumor delineations due to a great variety in tumor morphology with occasionally blurred tumor margins. Therefore, there is a need for further standardization.

Currently, the prediction of response to nCRT based on only qualitative (traditional subjective reading) and semi-quantitative (e.g., SUV parameters and total lesion glycolysis) baseline and restaging ^{18}F-FDG PET data seems to be insufficient. Multiple factors may contribute to an insufficient predictive power due to ^{18}F-FDG PET misinterpretation, including proper patient preparation and type of scanner. Besides, staging interpretation may be hindered by esophagitis (e.g., reflux induced, after endoscopic dilatation or radiotherapy), esophageal candidiasis, sarcoidosis, or low-glucose-metabolizing tumors (e.g., mucinous adenocarcinomas) [30]. Beyond ^{18}F-FDG PET radiomics, relevant information could be extracted from other functional imaging modalities such as DW-MRI or specific PET tracers. Since PET and MR images capture different intrinsic information about tumor biology, we strongly believe that such a multimodality approach would be able to optimize prediction of non-response to nCRT in esophageal cancer. Imaging information may also be complemented by genomic profiles of esophageal cancers, including data obtained from the Tumor Cancer Genome Atlas. Linking radiomic patterns directly to these genomic profiles, the so-called radiogenomics, may facilitate targeted treatment by radiomics-guided biopsy to specific sites identified with mutational burden or driven mutations.

A next logical step might be the implementation of deep learning algorithms such as convolutional neural networks. Advantages of convolutional neural networks include that features are automatically trained and no predefined handcrafted features are required in order to learn the relationship between input and outcome. Additionally, as tumor delineation is not essential, inconsistencies in delineation methods and intra- and interobserver variability can be reduced, increasing the accuracy. However, to ensure the generalization capability of such studies, even higher sample sizes are required due to the larger number of learnable parameters. This can be a practical limitation and can only be resolved by the standardization of used methods in collected studies, preferably in a multicenter prospective manner and international collaboration [31].

5. Conclusions

This is the first study to assess the value of ^{18}F-FDG PET radiomics combined with clinical features in the prediction of non-response to nCRT in esophageal cancer. Despite an extensive evaluation using various data dimension reduction techniques, classifiers, and training using cross-validation, we were only able to demonstrate a moderate discriminatory value for the constructed models. In the present study, the clinical impact of the best performing model, containing both clinical and ^{18}F-FDG PET-derived radiomic features, was not sufficient to predict non-response to nCRT in esophageal cancer.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12051070/s1, Table S1: displays the excluded vendor dependent and redundant features.

Author Contributions: Conceptualization, R.J.B., F.B.P., R.J.d.H., J.T.M.P. and J.B.H.; methodology, R.J.B., F.B.P., R.J.d.H., J.T.M.P., G.K.-U., A.R.V., R.B. and J.B.H.; software and analysis, R.J.B.; investigation, R.J.B., F.B.P. and J.B.H.; data curation, R.J.B., F.B.P. and J.B.H.; writing—original draft, R.J.B., F.B.P., G.K.-U., A.R.V., R.J.d.H., J.T.M.P. and J.B.H.; writing—review and editing, R.J.B., F.B.P., R.B., G.K.-U., A.R.V., R.J.d.H., J.T.M.P. and J.B.H.; supervision, R.J.d.H., J.T.M.P. and J.B.H.; project administration, R.J.B., J.T.M.P. and J.B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective study was conducted in accordance with the Dutch guidelines for retrospective studies and rules of the local institutional ethical board.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of this study, in accordance with the Dutch guidelines for retrospective studies and with the approval of the local institutional ethical board.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. van Hagen, P.; Hulshof, M.C.; van Lanschot, J.J.B.; Steyerberg, E.W.; Henegouwen, M.V.B.; Wijnhoven, B.P.L.; Richel, D.J.; Nieuwenhuijzen, G.A.P.; Hospers, G.A.P.; Bonenkamp, J.J.; et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N. Engl. J. Med.* **2012**, *366*, 2074–2084. [PubMed]
2. Kroese, T.E.; Goense, L.; van Hillegersberg, R.; De Keizer, B.; Mook, S.; Ruurda, J.P.; Van Rossum, P.S.N. Detection of distant interval metastases after neoadjuvant therapy for esophageal cancer with ^{18}F-FDG PET(/CT): A systematic review and meta-analysis. *Dis. Esophagus* **2018**, *31*, doy055. [CrossRef] [PubMed]
3. Chevrollier, G.S.; Giugliano, D.N.; Palazzo, F.; Keith, S.W.; Rosato, E.L.; Iii, N.R.E.; Berger, A.C. Patients with Non-response to Neoadjuvant Chemoradiation for Esophageal Cancer Have No Survival Advantage over Patients Undergoing Primary Esophagectomy. *J. Gastrointest. Surg.* **2020**, *24*, 288–298. [CrossRef] [PubMed]
4. Pribluda, A.; de la Cruz, C.C.; Jackson, E.L. Intratumoral heterogeneity: From diversity comes resistance. *Clin. Cancer Res.* **2015**, *21*, 2916–2923. [CrossRef]
5. Sengupta, D.; Pratx, G. Imaging metabolic heterogeneity in cancer. *Mol. Cancer* **2016**, *15*, 1–12. [CrossRef]
6. O'Connor, J.P.; Rose, C.J.; Waterton, J.C.; Carano, R.A.; Parker, G.J.; Jackson, A. Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome. *Clin. Cancer Res.* **2015**, *21*, 249–257. [CrossRef]
7. Lin, L.; Lin, D.C. Biological Significance of Tumor Heterogeneity in Esophageal Squamous Cell Carcinoma. *Cancers* **2019**, *11*, 1156. [CrossRef]
8. Gerashchenko, T.S.; Denisov, E.V.; Litviakov, N.V.; Zavyalova, M.V.; Vtorushin, S.V.; Tsyganov, M.M.; Perelmuter, V.M.; Cherdyntseva, N.V. Intratumor heterogeneity: Nature and biological significance. *Biochemistry* **2013**, *78*, 1201–1215. [CrossRef]
9. Limkin, E.J.; Sun, R.; Dercle, L.; Zacharaki, E.I.; Robert, C.; Reuzé, S.; Schernberg, A.; Paragios, N.; Deutsch, E.; Ferté, C. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann. Oncol.* **2017**, *28*, 1191–1206. [CrossRef]
10. Yang, Z.; He, B.; Zhuang, X.; Gao, X.; Wang, D.; Li, M.; Lin, Z.; Luo, R. CT-based radiomic signatures for prediction of pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant chemoradiotherapy. *J. Radiat. Res.* **2019**, *60*, 538–545. [CrossRef]

11. Hou, Z.; Ren, W.; Li, S.; Liu, J.; Sun, Y.; Yan, J.; Wan, S. Radiomic analysis in contrast-enhanced CT: Predict treatment response to chemoradiotherapy in esophageal carcinoma. *Oncotarget* **2017**, *8*, 104444–104454. [CrossRef] [PubMed]
12. Jin, X.; Zheng, X.; Chen, D.; Jin, J.; Zhu, G.; Deng, X.; Han, C.; Gong, C.; Zhou, Y.; Liu, C.; et al. Prediction of response after chemoradiation for esophageal cancer using a combination of dosimetry and CT radiomics. *Eur. Radiol.* **2019**, *29*, 6080–6088. [PubMed]
13. van Rossum, P.S.; Fried, D.V.; Zhang, L.; Hofstetter, W.L.; Van Vulpen, M.; Meijer, G.J.; Lin, S.H. 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*, 691–700. [CrossRef] [PubMed]
14. Beukinga, R.J.; Hulshoff, J.B.; Mul, V.E.M.; Noordzij, W.; Kats-Ugurlu, G.; Slart, R.H.J.A.; Plukker, J.T.M. Prediction of Response to Neoadjuvant Chemotherapy and Radiation Therapy with Baseline and Restaging ^{18}F-FDG PET Imaging Biomarkers in Patients with Esophageal Cancer. *Radiology* **2018**, *287*, 983–992. [CrossRef]
15. Tixier, F.; Le Rest, C.C.; Hatt, M.; Albarghach, N.; Pradier, O.; Metges, J.-P.; Corcos, L.; Visvikis, D. 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*, 369–378. [CrossRef] [PubMed]
16. Tan, S.; Kligerman, S.; Chen, W.; Lu, M.; Kim, G.; Feigenberg, S.; D'Souza, W.D.; Suntharalingam, M.; Lu, W. 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. [CrossRef]
17. Yip, S.S.; Coroller, T.P.; Sanford, N.N.; Mamon, H.; Aerts, H.J.; Berbeco, R.I. 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*, 72–82. [CrossRef]
18. Beukinga, R.J.; Hulshoff, J.B.; van Dijk, L.V.; Muijs, C.T.; Burgerhof, J.G.; Kats-Ugurlu, G.; Slart, R.H.; Slump, C.H.; Mul, V.E.; Plukker, J.T. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment ^{18}F-FDG PET/CT imaging. *J. Nucl. Med.* **2017**, *58*, 723–729. [CrossRef]
19. Nakajo, M.; Jinguji, M.; Nakabeppu, Y.; Nakajo, M.; Higashi, R.; Fukukura, Y.; Sasaki, K.; Uchikado, Y.; Natsugoe, S.; Yoshiura, T. 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*, 206–214. [CrossRef]
20. Bailly, C.; Bodet-Milin, C.; Bourgeois, M.; Gouard, S.; Ansquer, C.; Barbaud, M.; Sébille, J.-C.; Chérel, M.; Kraeber-Bodéré, F.; Carlier, T. Exploring Tumor Heterogeneity Using PET Imaging: The Big Picture. *Cancers* **2019**, *11*, 1282. [CrossRef]
21. Mandard, A.M.; Dalibard, F.; Mandard, J.C.; Marnay, J.; Henry-Amar, M.; Petiot, J.F.; Roussel, A.; Jacob, J.H.; Segol, P.; Samama, G.; et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer* **1994**, *73*, 2680–2686. [CrossRef]
22. Boellaard, R.; Delgado-Bolton, R.; Oyen, W.J.G.; Giammarile, F.; Tatsch, K.; Eschner, W.; Verzijlbergen, F.J.; Barrington, S.F.; Pike, L.C.; Weber, W.A.; et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: Version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* **2015**, *42*, 328–354. [CrossRef] [PubMed]
23. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
24. Pfaehler, E.; Van Sluis, J.; Merema, B.B.; van Ooijen, P.; Berendsen, R.C.; Van Velden, F.H.; Boellaard, R. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. *J. Nucl. Med.* **2020**, *61*, 469–476. [CrossRef]
25. Pfaehler, E.; Beukinga, R.J.; de Jong, J.R.; Slart, R.H.; Slump, C.H.; Dierckx, R.A.; Boellaard, R. Repeatability of ^{18}F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med. Phys.* **2019**, *46*, 665–678. [CrossRef] [PubMed]
26. Hatt, M.; Majdoub, M.; Vallières, M.; Tixier, F.; Le Rest, C.C.; Groheux, D.; Hindié, E.; Martineau, A.; Pradier, O.; Hustinx, R.; 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. [CrossRef] [PubMed]
27. Galavis, P.E.; 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*, 1012–1016. [CrossRef]
28. Yan, J.; Chu-Shern, J.L.; Loi, H.Y.; Khor, L.K.; Sinha, A.K.; Quek, S.T.; Tham, I.W.; Townsend, D.W. Impact of image reconstruction settings on texture features in ^{18}F-FDG PET. *J. Nucl. Med.* **2015**, *56*, 1667–1673. [CrossRef]
29. Whybra, P.; Parkinson, C.; Foley, K.; Staffurth, J.; Spezi, E. Assessing radiomic feature robustness to interpolation in ^{18}F-FDG PET imaging. *Sci. Rep.* **2019**, *9*, 9649. [CrossRef]
30. Flavell, R.R.; Naeger, D.M.; Aparici, C.M.; Hawkins, R.A.; Pampaloni, M.H.; Behr, S.C. Malignancies with Low Fluorodeoxyglucose Uptake at PET/CT: Pitfalls and Prognostic Importance: Resident and Fellow Education Feature. *Radiographics* **2016**, *36*, 293–294. [CrossRef]
31. Hatt, M.; Tixier, F.; Pierce, L.; Kinahan, P.E.; Le Rest, C.C.; Visvikis, D. Characterization of PET/CT images using texture analysis: The past, the present ... any future? *Eur. J. Nucl. Med. Mol. Imaging* **2017**, *44*, 151–165. [CrossRef] [PubMed]