Prediction of Lymph Node Metastases Using Pre-Treatment PET Radiomics of the Primary Tumour in Esophageal Adenocarcinoma: an External Validation Study

Chong Zhang, Zhenwei Shi, Petros Kalendralis, Phil Whybra, Craig Parkinson, Maaike Berbee, Emiliano Spezi, Ashley Roberts, Adam Christian, Wyn Lewis, Tom Crosby, Andre Dekker, Leonard Wee, Kieran G Foley

Abstract

Objectives: To improve clinical lymph node staging (cN-stage) in esophageal adenocarcinoma by developing and externally validating three prediction models; one with clinical variables only, one with positron emission tomography (PET) radiomics only, and a combined clinical and radiomics model.

Methods: Consecutive patients with fluorodeoxyglucose (FDG) avid tumours treated with neo-adjuvant therapy between 2010 and 2016 in two international centres (n=130 and n=60, respectively) were included. Four clinical variables (age, gender, clinical T-stage and tumour regression grade) and PET radiomics from the primary tumour were used for model development. Diagnostic accuracy, area under curve (AUC), discrimination and calibration were calculated for each model. The prognostic significance was also assessed.

Results: The incidence of lymph node metastases was 58% in both cohorts. The AUCs of the clinical, radiomics and combined models were 0.79, 0.69 and 0.82 in the developmental cohort, and 0.65, 0.63 and 0.69 in the external validation cohort, with good calibration demonstrated. The AUC of current cN-stage in development and validation cohorts was 0.60 and 0.66, respectively. For overall survival, the combined clinical and radiomics model achieved the best discrimination performance in the external validation cohort (X^2=6.08, df=1, p=0.01).

Conclusion: Accurate diagnosis of lymph node metastases is crucial for prognosis and guiding treatment decisions. Despite finding improved predictive performance in the development

cohort, the models using PET radiomics derived from the primary tumour were not fully replicated in an external validation cohort.

Advances in knowledge:

1. This international study attempted to externally validate a new prediction model for lymph node metastases using PET radiomics.

2. A model combining clinical variables and PET radiomics improved discrimination of lymph node metastases, but these results were not externally replicated.

Keywords: Neoplasms; Esophagus; Radiomics; Positron-Emission Tomography; Lymphatic Metastasis

Abbreviations

AJCC	American Joint Committee on Cancer
AUC	Area under curve
PET	Positron-emission tomography
FDG	Fluorodeoxyglucose
AUC	Area under curve
EC	Esophageal carcinoma
CT	Computed tomography
EUS	Endoscopic ultrasound
EUS-FNA	EUS fine needle aspiration
PET/CT	Positron-emission tomography-computed tomography
TNM	Tumour node metastasis
UICC	Union for International Cancer Control
MRI	Magnetic resonance imaging
NCT	Neo-adjuvant chemotherapy
NCRT	Neo-adjuvant chemoradiotherapy
TRG	Tumour regression grade
GTV	Gross tumour volume
ATLAAS	Automatic Decision Tree Learning Algorithm for Advanced Segmentation
CERR	Computational Environment for Radiological Research
SUV	Standardised uptake value
SPAARC	Spaarc Pipeline for Automated Analysis and Radiomics Computing
IBSI	Image Biomarker Standardization Initiative

RFE	Recursive feature elimination
LASSO	Least absolute shrinkage and selection operator
Rad-score	Radiomics score
MTV	Metabolic tumour volume
TLG	Total lesion glycolysis
pLNMS	Pathological lymph node metastasis
CI	Confidence interval
PPV	Positive predictive value
NPV	Negative predictive value
TRIPOD	Transparent Reporting of a multi-variable prediction model for Individual Prognosis or Diagnosis

Introduction

Esophageal carcinoma (EC) is the eleventh most common cancer and the sixth leading cause of cancer-associated death worldwide[1, 2] with adenocarcinoma being the most common histological cell type in many Western countries. The overall five-year survival rate of EC patients is 15%, with less than 40% of patients suitable for potentially curative therapy at presentation.[3] Importantly, lymph node metastases (LNMs) are a significant prognostic indicator of survival in EC.[4] Accurate knowledge of LNMs influences patient stratification, selection for radical therapy, treatment decision-making and planning.

Lymph nodes are assessed using computed tomography (CT), endoscopic ultrasound (EUS) and positron emission tomography with CT (PET/CT) as part of clinical Tumour Node Metastasis (TNM) staging.[5] Recent data suggests that the accuracy of lymph node staging with CT, EUS and PET/CT is poor (54.5%, 55.4% and 57.1%, respectively)[6]. The poor accuracy has been attributed to a high incidence of micro-metastases within morphologically normal-sized lymph nodes. These data are supported by a similar study which also demonstrated suboptimal N-staging accuracy (75.6%, 77.2% and 74.5%, respectively)[7]. This poor diagnostic accuracy translates to suboptimal patient selection and clinical outcomes because, despite aggressive treatment, the 2-year overall survival after neo-adjuvant chemotherapy and oesophagectomy ranges from 40-70%[8] with a 20% recurrence rate[9]. Thus, existing EC staging techniques are unlikely to detect small LNMs, so alternative biomarkers that improve diagnostic accuracy should be sought. The current difficulty in identifying LNMs is likely to be a contributor of poor treatment outcomes.

Advances in quantitative medical image data-mining techniques, broadly known as radiomics, enable the non-invasive decoding of tumour heterogeneity by translating medical images into abstract numerical features for analysis. In retrospective, single-centre studies, CT-derived

radiomics have enabled superior prediction of LNMs in colorectal cancer[10], bladder cancer[11], and esophageal cancer[12, 13]. These preoperative CT studies achieved satisfactory detection of LNMs for esophageal squamous cell carcinoma, reporting area under curve (AUC) statistics of 0.806 and 0.758 in development cohorts, and 0.771 and 0.773 in the validation cohorts, respectively.[12, 13] However, these results have not been reproduced in multi-centre or external validation studies, and therefore their clinical value remains unproven. Similar results have been reported using magnetic resonance imaging (MRI) radiomics, although MRI is often not routinely performed in the diagnostic pathway.[14]

PET radiomics have been significantly associated with overall survival[15], response to neoadjuvant therapy[16] and metastases[17], but the performance of PET radiomics extracted from the primary oesophageal adenocarcinoma to predict LNMs has not been tested. Increasing scientific evidence demonstrates that metastatic spread from the primary tumour is driven by biological changes in the underlying microenvironment of the primary tumour.[18] Generally, the additional value of radiomics extracted from small regions of interest, such as lymph nodes, over that of simple metrics such as volume is felt to be limited.[19] Accurate delineation is difficult and time-consuming which hinders the clinical utility of lymph node radiomics, unlike larger primary tumours which are more reliably outlined with less error.[20] Therefore, our study attempted to improve currently poor lymph node staging accuracy by extracting pre-treatment PET radiomics from the primary tumour to quantify its metastatic potential and predict the presence of LNMs following surgery.

In this study, we investigated the predictive value of PET radiomic features for LNMs by comparing three models: (1) a model based on clinical variables alone; (2) a model based on PET radiomics alone and (3) a combined model developed by clinical variables and PET radiomic features. The prognostic significance of developed LNM models was also assessed.

Materials & Methods

This study was a review board-approved Transparent Reporting of a multi-variable prediction model for Individual Prognosis or Diagnosis (TRIPOD) type 3 study (model development and external independent validation).[21] Research ethics committee approval was granted (reference 19/WA/0119).

Patients

To minimise selection bias, consecutive patients (n=190) with biopsy proven FDG-avid esophageal adenocarcinomas treated with neo-adjuvant therapy and surgery between 2010 and 2016 were included in this retrospective study. The development cohort (hereafter called "STAGE") comprised 130 patients receiving either surgery alone, neo-adjuvant chemo (NCT) or neo-adjuvant chemoradiotherapy (NCRT) followed by surgery in the *blinded.*[22] The external validation cohort (hereafter called "CROSS") comprised 60 patients who underwent NCRT at *blinded*. In both cohorts, the neo-adjuvant treatments were administered over a 12-week period. Patients with oesophageal stents in situ were excluded from the study. The PET/CT was performed prior to any treatment, but not repeated after neo-adjuvant therapy. This is common practice in many countries because the examination is expensive[23] and evidence for its cost-effectiveness in clinical practice is currently lacking. A proportion of these patients have previously been reported; 138 of 403 STAGE patients were reported in[15] and 46 of 60 CROSS patients in[24]. These prior articles developed a prognostic model for overall survival and validated the results in an external cohort. In the present study, we use standardised features to predict LNMs using radiomics from the primary tumour. The CROSS cohort were treated with the CROSS regimen[25-27] followed by resection of the esophagus after NCRT. **Figure 1** details the number of patients and exclusion criteria in each cohort.

Clinical Parameters

Routine clinical demographics were collected. Age was recorded at the time of diagnosis. Tumour location was recorded from a combination of endoscopic and radiological examinations. Radiological staging was assigned according to TNM 7th edition, which was used during the study period.[28] Tumour regression grade (TRG) was defined using the Mandard classification.[29] The primary outcome was LNM status, determined by histopathological examination. Overall survival was defined in months from the date of diagnosis until date of death or last follow-up.

Radiomics Feature Extraction and Tumour Segmentation

To reduce interobserver variability, esophageal primary gross tumour volumes (GTVs) were systematically delineated on PET images using "Automatic Decision Tree Learning Algorithm for Advanced Segmentation" (ATLAAS).[30] ATLAAS was implemented in MATLAB (The Mathworks, Natick, USA) as a plug-in to the Computational Environment for Radiological Research (CERR).[31] PET images were re-sampled into 0.5 standardised uptake value (SUV) equally sized bins, which has been recommended.[32] Voxel size was interpolated to 2 x 2 x 2 mm prior to feature extraction. In total, 154 radiomic features were extracted from the GTV using the Spaarc Pipeline for Automated Analysis and Radiomics Computing (SPAARC).[33] SPAARC radiomic features comply with the Image Biomarker Standardization Initiative (IBSI).[34] Different scanners and imaging protocols were used across the two centres. Radiomic features could be changed significantly as a function of scanner, image acquisition or reconstruction settings, hence we performed the post-reconstruction Combat harmonization method[35] to harmonize features extracted from images acquired across different centres. Detailed information about image acquisition and parameter settings are included in the supplementary materials.

Feature Selection and Prediction Model Development

Model development was performed blinded to pathological assessment. Missing data were excluded from model development. A flowchart describing feature selection and model development in the STAGE cohort is shown in **Figure 2**. Recursive Feature Elimination (RFE) and a Least Absolute Shrinkage and Selection Operator (LASSO) method were used to select the optimal clinical feature combination (selected from age, gender, tumour location, histological cell type, clinical T-stage, type of neo-adjuvant treatment and tumour regression grade (Mandard score)), using the AUC measurement from the receiver operator curve (ROC). Clinical N-stage (cN-stage) was collected for comparative analysis and to evaluate the baseline staging accuracy in each cohort, but was not included in the multivariable models to avoid multi-collinearity and prevent influencing the model by using potentially inaccurate data.

For radiomics features, pair-wise Pearson correlation of radiomic features was calculated and the threshold was set to 0.85. The cut-off logic assessed the mean absolute correlation of each radiomic feature and removed the feature with the largest mean absolute correlation. A non-parametric Kolmogorov-Smirnov test statistic was calculated for each feature between resected node positive (pN+) and negative (pN0) outcomes and only features with a p-value <0.05 were retained to ensure the two classes were significantly distinguished. A LASSO method was used to tabulate the most frequently-selected radiomic features over 500 repetitions of internally separating STAGE into training (70%) and validation (30%) subsets. The best lambda of LASSO was automatically selected in each repetition based on AUC. Finally, we used RFE with 5-fold cross validation to search for combinations of features with non-zero frequency, to find an optimal combination by its AUC.

Feature value normalization before RFE was performed using the mean and standard deviation of selected features in STAGE. Prediction models were developed using multivariable logistic regression in the STAGE cohort and a radiomics score (Rad-score) for each patient was then computed using the coefficients weighted by regression model. A combined model was

developed using the selected clinical variables and the Rad-score. External validation in CROSS was performed using the same data transformations that were applied in STAGE.

Statistical Analysis

Statistical analyses were performed in R (v3.30). Clinical demographic differences between STAGE and CROSS were examined by two-sided t-test (continuous variables) or chi-square/fisher test (categorical variables). Estimates of 95% confidence intervals were derived from 2000 stratified bootstrap replicates. Appropriate calibration of the models were assessed using calibration plots and Hosmer-Lemeshow test statistics.

Prognostic significance was explored by entering the selected clinical variables and radiomics features to a Cox proportional hazards model with censoring. We computed the Harrell concordance index and performed log-rank tests of significance for the survival models. To ensure the higher-order radiomics variables were not just surrogates for simple tumour characteristics, we also compared the concordance indices and log-rank tests against primary metabolic tumour volume (MTV) and total lesion glycolysis (TLG).

Results

The patient characteristics of the STAGE and CROSS cohorts are detailed in **Table 1**. The incidences of pathological lymph node metastasis (pLNMs) were 58% (75/130) and 58% (35/60) in STAGE and CROSS, respectively. In STAGE, the majority (62.3%) had NCT whereas all CROSS patients had NCRT. The cohorts differed significantly for cN staging and tumour location. mean follow-up times were 25.6 months (95% confidence interval (CI): 22.7-28.4) in the STAGE and 28.5 months (95% CI: 23.6-33.4) in the CROSS cohorts, respectively. A log-rank hypothesis test showed no significant difference in actuarial survival between the STAGE and CROSS cohorts (p=0.237).

Four clinical features; age, clinical T-stage, neo-adjuvant therapy and Mandard classification, were included in the multivariable model after applying RFE method optimized for AUC. These features were within the top four most-frequently selected directly through LASSO during 500 random splits of STAGE. This multivariable clinical model yielded mean AUCs of 0.79 (95% CI: 0.71-0.88)) in STAGE and 0.65 (95% CI: 0.50-0.78) in CROSS. In the same cohorts, a cN-based model resulted in mean AUCs 0.60 (95% CI: 0.52-0.69) and 0.66 (95% CI: 0.55-0.78), respectively.

Nine radiomic features were selected for a radiomics-based model, but resulted in lower mean AUCs of 0.69 (95% CI: 0.59-0.77) in STAGE and 0.63 (95% CI: 0.47-0.77) in CROSS. A combined clinical and radiomics-based model yielded mean AUCs of 0.82 (95% CI: 0.74-0.89) and 0.69 (95% CI: 0.54-0.82) in these cohorts, respectively. In the validation cohort, there was no statistically significant difference in AUC performance across the three models. **Figure 3** shows the ROC plots of the above models with their respective mean AUCs. Results of AUC, accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are reported in **Table 2**. cN-stage results for each cohort were calculated and are included in **Table 2** for comparison with each of the three models.

The calibration plots of the models in both cohorts are shown in **Supplementary Figures 1**. The Hosmer-Lemeshow test indicated that the combined model was well calibrated in for development ($p=0.11$) and validation ($p=0.47$), although calibration was suboptimal for clinical only ($p=0.02$) and radiomics only models ($p=0.01$) in the development cohort.

In univarable analysis, the AUC for MTV to predict LNMs was 0.58 (95% CI: 0.48-0.66) and 0.60 (95% CI: 0.47-0.72) in STAGE and CROSS cohorts, respectively. For TLG, the AUC was 0.58 (95% CI: 0.48-0.66) and 0.58 (95% CI: 0.45-0.71), respectively.

Results of survival analysis are tabulated in **Supplementary Table 1** and **2 (ST-1, 2)** and Kaplan-Meier curves are given in **Supplementary Figure 2, 3, 4 (SF-2, 3, 4)**. In both STAGE and CROSS cohorts, the true pathological lymph node status (X^2 13.76, df 1, $p < 0.001$, and X^2 4.36, df 1, $p = 0.04$, respectively, **ST-1&2**) was significantly associated with overall survival. In addition, the combined clinical and radiomics model was also significantly associated with overall survival in the external CROSS cohort (X^2 6.08, df 1, $p = 0.01$, **Supplementary Figure 4**) and performed better than the other developed models. Finally, **Supplementary Figure 5, 6, 7 (SF-5, 6, 7)** show the performance of the three types of models in the development cohort according to the subgroups divided by treatment types.

Discussion

In this study, we developed and externally validated three prediction models; a model using clinical variables only, PET radiomics only and a combined clinical-radiomics model. A combined clinical and radiomics model developed in the STAGE cohort showed potentially improved diagnostic performance compared with current cN-stage results but this was not replicated in the external validation CROSS cohort.

In terms of prognostic significance, a combined clinical-radiomics model demonstrated good discrimination between patient groups in the external cohort but this was not the case in the development cohort. The external cohort included patients recruited into the CROSS trial[27], in which the pathological stage following NACRT (ypTN+ stage) was significantly associated with overall survival. There were significant differences in cN-stage status between cohorts, with a

higher proportion of patients having cN+ disease in the external CROSS cohort. This was reflected in the sensitivity and specificity results obtained. The sensitivity was reduced in the STAGE compared to CROSS cohorts with the opposite true for specificity, indicating that radiologists were more likely to 'under-stage' disease in STAGE compared to radiologists in CROSS. The variability in staging classification maybe explained by reporting practice differences between the two countries.

Failure to replicate models is a common finding in external validation studies. The lack of full validation may be attributable to the relatively small sample size of the external validation cohort, inter-scanner differences such as varying slice thickness, voxel size and acquisition parameters, and inter-patient differences such as time from injection to imaging. However, PET radiomics have been shown to have potential clinical value in EC when incorporated into a prognostic model.[15]

Initially, the results showed that PET radiomics derived from the primary tumour volume may have added predictive value for LNM detection, but this effect was not replicated. Common concerns are that firstly, clinical PET images have a spatial resolution too large for radiomic analysis and secondly, higher order radiomic variables are surrogates of simple MTV.[36] However, simpler PET metrics such as MTV and TLG were excluded as potential confounders, through a detailed process of radiomic feature selection. Furthermore, MTV and TLG had no predictive value for LNMs or overall survival. Despite comprehensive clinical and radiological data, we failed to show that a radiomics signature was significantly superior to either cN-stage or the clinical multivariable model for predicting LNMs in esophageal adenocarcinoma.

Our results add evidence that current cN-staging accuracy remains poor.[6, 7] The Union for International Cancer Control (UICC) Tumour Node Metastasis (TNM) classification is largely reliant on anatomical definitions.[28] CT has sensitivity for LNMs as low as 18%[37] because it relies

on morphology, and cannot differentiate between occult malignant metastases and normal-sized lymph nodes. It is now thought that exosomes are excreted by aggressive primary tumours into the bloodstream. This circulating tumour DNA (ctDNA) seeds in lymph nodes, giving rise to synchronous micro-metastases.[38] Patients with esophageal adenocarcinomas commonly present with LNMs due to lack of esophageal serosa.[39] EC staging must become more accurate and better at risk-stratifying patients.

Commonly, treatment decisions often hinge on the accurate diagnosis of a lymph node. (**Figure 4**) For example, equivocal lymph nodes located away from the primary tumour may be considered un-resectable or be outside of the maximum radiotherapy field possible.[40] Often, tissue confirmation is attempted with EUS fine needle aspiration (EUS-FNA) but on occasions where the aspirate is normal or insufficient, concerns over under-sampling exist. Additional predictive information would add confidence to this important treatment decision. Similarly, improved diagnostic accuracy of non-regional lymph nodes in the abdomen that are inaccessible by FNA or core biopsy may prevent a harmful major resection that is unlikely to yield long-term survival gain. Finally, in the case of T1-T2 N0 tumours, the decision to proceed directly to surgery is standard practice. However, the risk of pLNMs is 45–75% for T2 tumours and 80–85% for >= T3 tumours.[41] Administration of neo-adjuvant therapy would be preferable in these scenarios. Non-invasive imaging biomarkers that suggest that the primary tumour has high metastatic potential would guide the clinical decision towards neo-adjuvant treatment prior to surgery. Further research that focusses on this disease stage group is warranted but would require a large sample size to adequately power such a study.

Strengths of study

External validation studies are rarely performed, particularly in the field of radiomics. To eliminate the inter-observer variability of delineation, we used a standardised auto-segmentation

approach (ATLAAS) to outline the tumour on PET images. Standardised biomarkers (IBSI) were also used. These reproducible methods allow further validation in different centres. Furthermore, a reproducible radiomic feature selection method was employed. In principle, as fewer important features are used in the model, the chance of model overfitting reduces. A common criticism of radiomics studies are that large numbers of predictor variables are used for a relatively small cohort size.[42] Therefore, we performed a relatively strict feature reduction approach to maximise the event per variable ratio.

Limitations

The main limitation was the lack of PET radiomics following neo-adjuvant therapy. As discussed, PET/CT is not repeated prior to surgery in many countries due to limited evidence about its cost-effectiveness, therefore quantifying the change in radiomics over time is not possible. As a result, lymph node response to neo-adjuvant treatment on PET cannot be assessed in this study and subsequently there is indirect comparison between baseline radiological features and the final pathological lymph node evaluation following surgery. This does, however, reflect real-world practice in many institutions around the world, because PET-CT re-staging after neoadjuvant therapy but before surgery is generally not performed. Arguably, the greatest advance would be the knowledge at baseline that lymph node metastases are present so that treatment could be tailored accordingly. Repeat analysis after neoadjuvant therapy has been completed offers little change in management, because all oncological treatment has been given and the majority of patients will proceed to surgery irrespective of the outcome. However, 63% and 40% of patients had a TRG of 4 or 5 in the STAGE and CROSS cohorts, repectively, indicating that a substantial proportion of patient had little or no response to neo-adjuvant therapy. This provides some reassurance that an indirect comparison provides some meaningful data. The differences in TRG rates can be explained by the differences in treatment between the two cohorts, with CROSS patients receiving NCRT but

the majority of STAGE receiving NCT. Secondly, only primary tumours were analysed. Integrated radiomics analysis of the primary tumour and individual lymph nodes may potentially provide more prediction information[43] but this process is more time-consuming and unlikely to be adopted into busy clinical practice. In addition, only FDG-avid adenocarcinomas were included in the study. Analyses of squamous cell carcinoma and non FDG-avid tumours would be equally valuable.

Conclusions

Accurate diagnosis of LNMs is crucial for predicting prognosis and guiding treatment decisions in esophageal adenocarcinoma, but radiological cN-staging is currently suboptimal. Despite obtaining signal for improved prediction in a development cohort, this study showed that models using clinical variables and PET radiomics derived from the primary tumour were not fully replicated in an external validation cohort from an international centre. We plan to further validate and confirm these findings in larger external cohorts. New techniques for improving the diagnostic accuracy of LNMs are required.

References

1. Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncology. 2017;3(4):524-48.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a Cancer Journal for Clinicians. 2015;65(2):87-108.
3. Cancer Research UK. Oesophageal cancer statistics. 2020 [accessed 2020 29, April]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#heading-Two.
4. Kayani B, Zacharakis E, Ahmed K, Hanna G. Lymph node metastases and prognosis in oesophageal carcinoma–a systematic review. European Journal of Surgical Oncology (EJSO). 2011;37(9):747-53.
5. Allum W, Griffin S, Watson A, Colin-Jones D. Guidelines for the management of oesophageal and gastric cancer. Gut. 2002;50(suppl 5):v1-v23.
6. Foley K, Christian A, Fielding P, Lewis W, Roberts S. Accuracy of contemporary oesophageal cancer lymph node staging with radiological-pathological correlation. Clinical Radiology. 2017;72(8):e691-e97.
7. Bunting D, Bracey T, Fox B, Berrisford R, Wheatley T, Sanders G. Loco-regional staging accuracy in oesophageal cancer—How good are we in the modern era? European Journal of Radiology. 2017;97:71-75.
8. Allum W, Stenning S, Bancewicz J, Clark P, Langley R. Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in esophageal cancer. Journal of Clinical Oncology. 2009;27:5062-67.
9. Turkington R, Knight L, Blayney J, Secrier M, Douglas R, Parkes E, et al. Immune activation by DNA damage predicts response to chemotherapy and survival in oesophageal adenocarcinoma. Gut. 2019;68(11):1918-27.
10. Huang Y-Q, Liang C-H, He L, Tian J, Liang C-S, Chen X, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. Journal of Clinical Oncology. 2016;34(18):2157-64.
11. Wu S, Zheng J, Li Y, Yu H, Shi S, Xie W, et al. A radiomics nomogram for the preoperative prediction of lymph node metastasis in bladder cancer. Clinical Cancer Research. 2017;23(22):6904-11.
12. Shen C, Liu Z, Wang Z, Guo J, Zhang H, Wang Y, et al. Building CT radiomics based nomogram for preoperative esophageal cancer patients lymph node metastasis prediction. 2018;11(3):815-24.
13. Tan X, Ma Z, Yan L, Ye W, Liu Z, Liang C. Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma. European Radiology. 2019;29(1):392-400.
14. Qu J, Shen C, Qin J, Wang Z, Liu Z, Guo J, et al. The MR radiomic signature can predict preoperative lymph node metastasis in patients with esophageal cancer. European Radiology. 2019;29(2):906-14.
15. Foley KG, Hills RK, Berthon B, Marshall C, Parkinson C, Lewis WG, et al. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. European Radiology. 2018;28(1):428-36.

16. van Rossum PS, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, et al. The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. Journal of Nuclear Medicine. 2016;57(5):691-700.
17. Dong X, Xing L, Wu P, Fu Z, Wan H, Li D, et al. Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumor stage. Nuclear Medicine Communications. 2013;34(1):40-46.
18. Walker RC, Underwood TJ. Molecular pathways in the development and treatment of oesophageal cancer. Best Practice & Research Clinical Gastroenterology. 2018;36-37:9-15.
19. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Frontiers in Oncology. 2016;6:71.
20. Hatt M, Lee JA, Schmidtlein CR, Naqa IE, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Medical physics. 2017;44(6):e1-e42.
21. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Medicine. 2015;13(1):1.
22. Foley KG, Fielding P, Lewis WG, Karran A, Chan D, Blake P, et al. Prognostic significance of novel 18F-FDG PET/CT defined tumour variables in patients with oesophageal cancer. European Journal of Radiology. 2014;83(7):1069-73.
23. Schreurs LM, Janssens A, Groen H, Fockens P, van Dullemen HM, van Berge Henegouwen MI, et al. Value of EUS in determining curative resectability in reference to CT and FDG-PET: the optimal sequence in preoperative staging of esophageal cancer? Annals of Surgical Oncology. 2016;23(5):1021-28.
24. Foley KG, Shi Z, Whybra P, Kalendralis P, Larue R, Berbee M, et al. External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiotherapy and Oncology. 2019;133:205-12.
25. Noordman BJ, Verdam MG, Lagarde S, Hulshof M, Van Hagen P, van Berge Henegouwen MI, et al. Effect of neoadjuvant chemoradiotherapy on health-related quality of life in esophageal or junctional cancer: results from the randomized CROSS trial. Journal of Clinical Oncology. 2018;36(3):268-75.
26. Shapiro J, Van Lanschot JJB, Hulshof MC, van Hagen P, van Berge Henegouwen MI, Wijnhoven BP, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncology. 2015;16(9):1090-98.
27. van Hagen P, Hulshof M, Van Lanschot J, Steyerberg E, Henegouwen MVB, Wijnhoven B, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. New England Journal of Medicine. 2012;366(22):2074-84.
28. Sobin LH, Gospodarowicz MK, Wittekind C. TNM classification of Malignant Tumours (7th edition). New York: John Wiley & Sons, 2011.
29. Mandard AM, Dalibard F, Mandard JC, Marnay J, Henry-Amar M, Petiot JF, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. Cancer. 1994;73(11):2680-86.
30. Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. Physics in Medicine & Biology. 2016;61(13):4855.

31. Apte AP, Iyer A, Crispin-Ortuzar M, Pandya R, van Dijk LV, Spezi E, et al. Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. Medical Physics. 2018;45(8), 3713-3720.
32. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific Reports. 2015;5:11075.
33. Whybra P, Parkinson C, Foley K, Staffurth J, Spezi E. Assessing radiomic feature robustness to interpolation in 18 F-FDG PET imaging. Scientific Reports. 2019;9(1):9649.
34. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative-feature definitions. Radiology. 2020;295(2):191145.
35. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
36. Cook GJ, Siddique M, Taylor BP, Yip C, Chicklore S, Goh V. Radiomics in PET: principles and applications. Clinical and Translational Imaging. 2014;2(3):269-76.
37. Choi J, Lee K, Shim Y, Lee K, Kim J, Kim S. Improved detection of individual nodal involvement in squamous cell carcinoma of the esophagus by FDG PET. Journal of Nuclear Medicine. 2000;41(5):808-15.
38. Becker A, Thakur B, Weiss J, Kim H, Peinado H, Lyden D. Extracellular vesicles in cancer: cell-to-cell mediators of metastasis. Cancer Cell. 2016;30(6):836-48.
39. Rustgi AK. Gastrointestinal cancers: a companion to Sleisenger and Fordtran's Gastrointestinal and Liver Disease. London: WB Saunders, 2003.
40. Foley KG, Morgan C, Roberts SA, Crosby T. Impact of positron emission tomography and endoscopic ultrasound length of disease difference on treatment planning in patients with oesophageal cancer. Clinical Oncology. 2017;29(11):760-66.
41. Hölscher A, Bollschweiler E, Bumm R, Bartels H, Höfler H, Siewert J. Prognostic factors of resected adenocarcinoma of the esophagus. Surgery. 1995;118(5):845-55.
42. Chalkidou A, O'Doherty M, Marsden P. False discovery rates in PET and CT studies with texture features: a systematic review. Plos One. 2015;10(5):e0124165.
43. Coroller T, Agrawal V, Huynh E, Narayan V, Lee S, Mak R. Radiomic-based pathological response prediction from primary tumors and lymph nodes in NSCLC. Journal of Thoracic Oncology. 2017;12(3):467-76.

Table and Figure legends:

Table 1: Patient Characteristics in STAGE and CROSS Cohort.

Characteristic; Frequency (%)	STAGE Development Cohort (n=130)	CROSS Validation Cohort (n= 60)	p-value
Tumour type			$P = 1.00
Adenocarcinoma	130 (100%)	60 (100%)	#P = 0.63
Age mean ± SD, years	64.33 ± 9.54	63.15 ± 8.68	
Gender			$P = 0.38
Male	111 (85.4%)	54 (90.0%)	
Female	19 (14.6%)	6 (10.0%)	
Tumor location			+P = 0.0059
Distal third esophagus	45 (34.6%)	34 (56.6%)	
Mid third esophagus	7 (5.4%)	1 (1.7%)	
Esophagastric junction	78 (60%)	24 (41.7%)	
Pathological LNMs			+P = 0.94
Negative	55 (42.3 %)	25 (41.7%)	
Positive	75 (57.7%)	35 (58.3%	+P = 0.12
Clinical T stage			
T1	5 (3.8%)	0 (0.0)	
T2	14 (10.7%)	12 (20%)	
T3	101 (77.7%)	46 (76.7%)	
T4a	10 (7.8%)	2 (3.3%)	+P < 0.001
Clinical N stage			
N0	60 (46.2%)	15 (25.0%)	
N1	50 (38.5%)	17 (28.3%)	
N2	13 (10.0%)	15 (25.0%)	
N3	7 (5.3%)	13 (21.7%)	
Stage Groups			+P = 0.16
Stage 1	17 (13.1%)	4 (6.7%)	
Stage 2	43 (33.1%)	15 (25.0%)	
Stage 3	70 (53.8%)	41 (68.3%)	
TRG Score			+P < 0.08
1	12 (9.2%)	11 (18.3%)	
2	12 (9.2%)	11 (18.3%)	
3	13 (10.0%)	14 (23.3%)	
4	37 (28.5%)	16 (26.7%)	
5	26 (20.0%)	8 (13.4%)	
Not applicable	30 (23.1%)	0 (0.0)	
Neo-adjuvant therapy			+P < 0.001
NCRT	19 (14.6%)	60 (100%)	
NCT	81 (62.3%)	0 (0.0)	
Surgery alone	30 (23.1%)	0 (0.0)	
Overall survival			NA

Alive	77 (59.2%)	28 (46.7%)	
Dead	53 (40.8%)	32 (53.3%)	
Radiomics score, mean ± SD	0.49 ± 1.80	0.72 ± 2.78	#P =0.56

LNMs lymph node metastases; NCT neo-adjuvant chemotherapy; NCRT neo-adjuvant chemoradiotherapy; $ chi-square test; # t-test; + fisher test; NA not applicable

Table 2: The statistic comparison of clinical N-staging, clinical model, radiomics-based model, and combined model in the development and external validation cohorts.

	Development Cohort				External Validation Cohort			
	Clinical N-stage	**Clinical model**	**Radiomics model**	**Combined model**	**Clinical N-stage**	**Clinical model**	**Radiomics model**	**Combined model**
Incidence	58%	58%	58%	58%	58%	58%	58%	58%
AUC	0.60 (95% CI: 0.52-0.69)	0.79 (95% CI: 0.71-0.88)	0.69 (95% CI: 0.59-0.77)	0.82 (95% CI: 0.74-0.89)	0.66 (95% CI: 0.55-0.78)	0.65 (95% CI: 0.50-0.78)	0.63 (95% CI: 0.47-0.77)	0.69 (95% CI: 0.54-0.82)
Accuracy	0.61 (95% CI: 0.52-0.69)	0.79 (95% CI: 0.70-0.85)	0.66 (95% CI: 0.57-0.74)	0.76 (95% CI: 0.71- 0.79)	0.70 (95% CI: 0.57-0.81)	0.57 (95% CI: 0.43-0.69)	0.65 (95% CI: 0.52-0.75)	0.65 (95% CI: 0.51-0.76)
Sensitivity	0.634 (95% CI: 0.60-0.65)	0.88 (95% CI: 0.83-0.90)	0.77 (95% CI: 0.83-0.90)	0.81 (95% CI: 0.79-0.92)	0.89 (95% CI: 0.86-0.90)	0.52 (95% CI: 0.44-0.59)	0.74 (95% CI: 0.65-0.77)	0.63 (95% CI: 0.56-0.67)
Specificity	0.58 (95% CI: 0.57-0.60)	0.65 (95% CI: 0.62-0.69)	0.49 (95% CI: 0.47-0.50)	0.66 (95% CI: 0.62-0.69)	0.44 (95% CI: 0.41-0.46)	0.64 (95% CI: 0.44-0.69)	0.48 (95% CI: 0.46-0.54)	0.64 (95% CI: 0.59-0.70)
PPV	0.67 (95% CI: 0.64-0.69)	0.78 (95% CI: 0.74-0.79)	0.68 (95% CI: 0.64-0.71)	0.76 (95% CI: 0.72-0.78)	0.69 (95% CI: 0.67-0.73)	0.667 (95% CI: 0.56-0.69)	0.67 (95% CI: 0.63-0.70)	0.71 (95% CI: 0.69-0.74)
NPV	0.53 (95% CI: 0.51-0.57)	0.80 (95% CI: 0.76-0.82)	0.61 (95% CI: 0.56-0.70)	0.72 (95% CI: 0.70-0.90)	0.73 (95% CI: 0.70-0.756)	0.49 (95% CI: 0.43-0.59)	0.57 (95% CI: 0.52-0.62)	0.55 (95% CI: 0.53-0.66)

AUC area under curve; CI confidence interval; PPV positive predictive value; NPV negative predictive value.

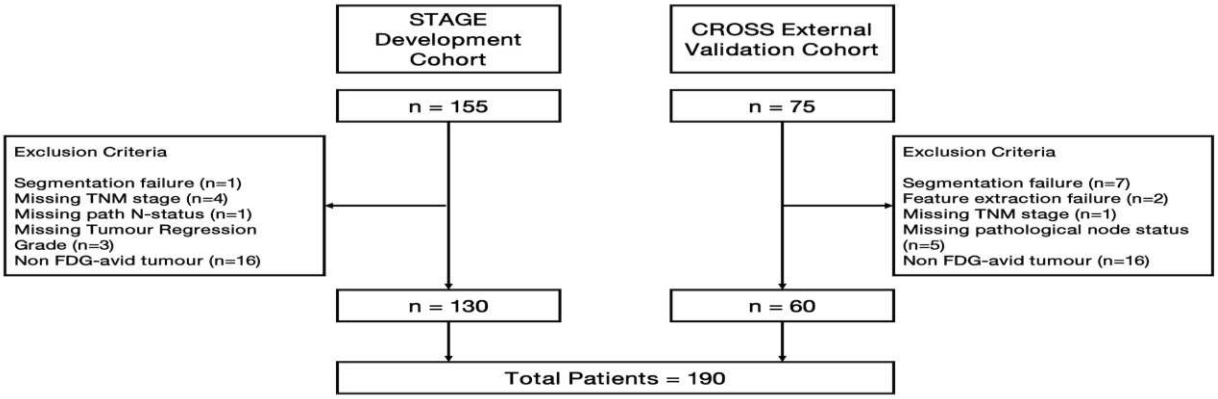


Figure 1: Study flowchart describing the numbers of patients in each cohort and reasons for exclusions from the CROSS cohort.

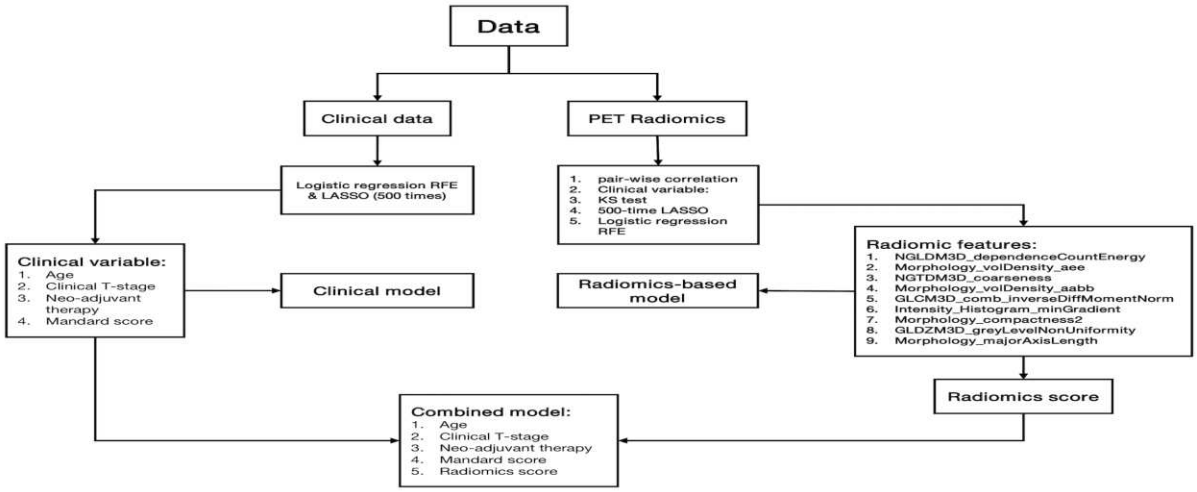
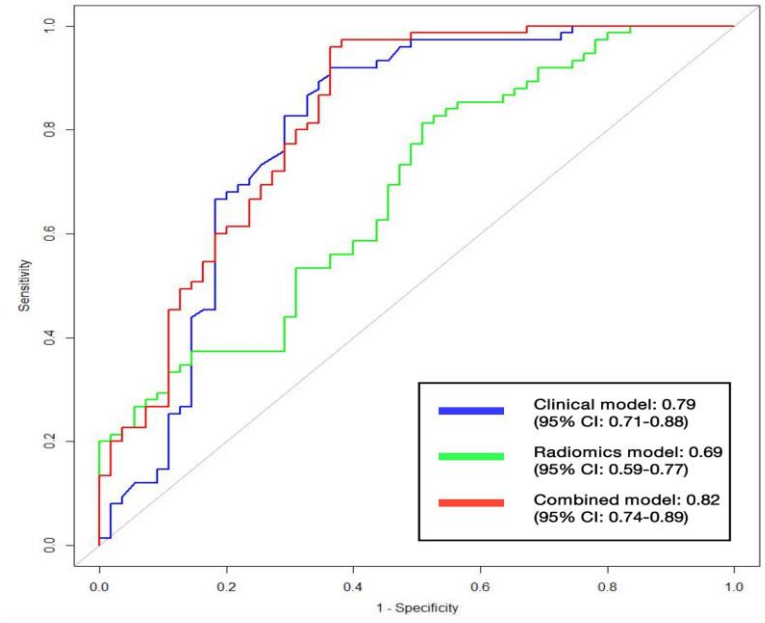
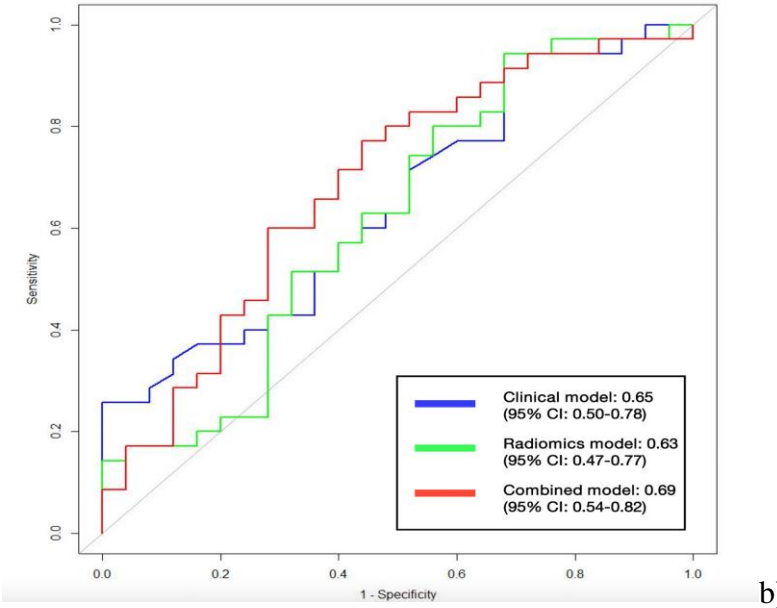


Figure 2: The flowchart of the used feature selection and model development approaches.



a)



b)

Figure 3: ROC plots of clinical model (blue line), radiomics-based model (green line), and combined model (red line) in the (a) development and (b) external validation cohorts. The results show the combined model achieved the best performance with AUCs 0.82 (95% CI: 0.74-0.90) and 0.71 (95% CI: 0.59-0.83) in the development and external validation cohorts. 95% CI was computed with 2000 times bootstrapping. CI: confidence interval.

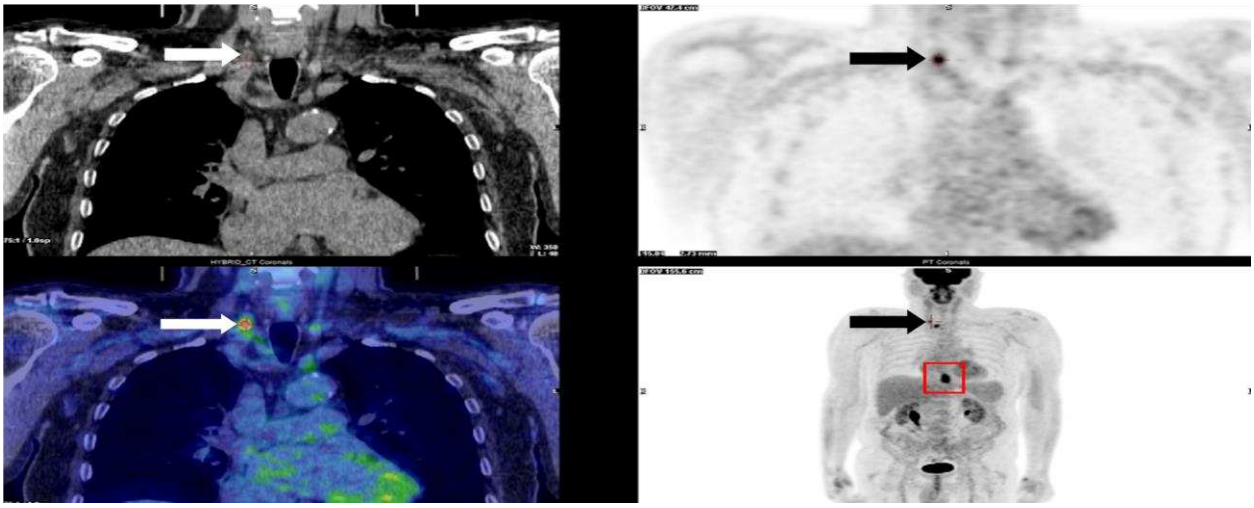


Figure 4: Clockwise from top left; a non-contrast CT, a magnified maximum intensity projection (MIP), a whole body MIP and fused PET/CT image of a patient with a distal esophageal tumour (Red box) and two lnms; one supra-clavicular and one high right para-esophageal (black and white arrows).