External Validation of a Prognostic Model Incorporating Quantitative PET Image Features in Esophageal Cancer

KG Foley[1], Z Shi[2], P Whybra[3], P Kalendralis[2], R Larue[2], M Berbee[2], Sosef MN[4], C Parkinson[3], J Staffurth[1, 5], TDL Crosby[5], SA Roberts[6], A Dekker[2], L Wee[2] and E Spezi[3]

1. Division of Cancer & Genetics, School of Medicine, Cardiff University, UK
2. Department of Radiation Oncology (MAASTRO Clinic), GROW – School for Oncology and Development Biology, Maastricht University Medical Centre, The Netherlands
3. School of Engineering, Cardiff University, UK
4. Zuyderland Medisch Centrum, Heerlen-Sittard-Geleen, The Netherlands
5. Velindre Cancer Centre, Cardiff, UK
6. Department of Clinical Radiology, University Hospital of Wales, Cardiff, UK

Corresponding Author
Dr KG Foley
Division of Cancer & Genetics, School of Medicine, Cardiff University, CF14 4XN
Tel: +447795152790
Fax: +442920743029
E-mail: foleykg@cardiff.ac.uk

Abstract

Aim

Enhanced prognostic models are required to improve risk stratification of patients with esophageal cancer so treatment decisions can be optimised. The primary aim was to externally validate a published prognostic model incorporating PET image features. Transferability of the model was compared using only clinical variables.

Methods

This was a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 study. The model was validated against patients treated with neoadjuvant chemoradiotherapy according to the Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for esophageal or junctional cancer (CROSS) trial regimen using pre- and post-harmonised image features. The Kaplan-Meier method with log-rank significance tests assessed risk strata discrimination. A Cox proportional hazards model assessed model calibration. Primary outcome was overall survival (OS).

Results

Between 2010 and 2015, 449 patients were included in the development (n=302), internal validation (n=101) and external validation (n=46) cohorts. No statistically significant difference in OS between patient quartiles was demonstrated in prognostic models incorporating PET image features (X^2=1.42, df=3, p=0.70) or exclusively clinical variables (age, disease stage and treatment; X^2=1.19, df=3, p=0.75). The calibration slope β of both models was not significantly different from unity (p=0.29 and 0.29, respectively). Risk groups defined using only clinical variables suggested differences in OS, although these were not statistically significant (X^2=0.71, df=2, p=0.70).

Conclusion

The prognostic model did not enable significant discrimination between the validation risk groups, but a second model with exclusively clinical variables suggested some transferable prognostic ability. PET harmonisation did not significantly change the results of model validation.

Keywords: esophageal cancer; positron-emission tomography; radiomics; survival; prognosis

List of Abbreviations

LNMs	lymph node metastases
PET	positron-emission tomography
NACRT	neo-adjuvant chemoradiotherapy
CROSS	Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer
TRIPOD	Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis
GI	gastrointestinal
MDT	multi-disciplinary team
CaNISC	Cancer Network Information System
ATLAAS	Automatic Tree-based Learning Algorithm for Advanced Segmentation
CI	confidence interval
TLG	tumour lesion glycolysis
OS	overall survival
IBSI	International Biomarker Standardisation Initiative

Introduction

The prognosis of patients with esophageal cancer is poor with overall 5-year survival approximately 15%. [1] Esophageal cancer is the eighth most common malignancy worldwide, accounting for around 400,000 deaths each year. [2]

Treatment strategies of patients with esophageal cancer are currently informed by radiological staging. Accurate staging is vital to inform clinicians of the likely prognosis of each patient and to appropriately risk stratify patients, ensuring the best individual management plan is decided upon. However, the stagnation in survival rate over recent decades suggests that staging accuracy, treatment selection and prognosis could be much improved. For example, lymph node metastases (LNMs) are one of the major prognostic indicators in esophageal cancer, but there is evidence that regional lymph node staging (N-stage) is presently suboptimal. [3, 4] Therefore, enhanced staging methods are required to improve prognostication and subsequent risk stratification of patients.

Esophageal cancer is typically confirmed by a small-sample biopsy taken during endoscopic examination. Despite advances in genomics, no molecular prognostic markers are currently in routine clinical use. [5] It has been proposed that additional tumour phenotype information may be derived by quantitative analysis of Positron Emission Tomography (PET) scans. [6] "Radiomics" broadly refers to automated, computerised and high-throughput extraction of quantitative image markers (features) from a large corpus of radiological images. [7] Radiomics features typically include histogram metrics (e.g. mean and maximum), shape descriptors (e.g. longest axis length and compactness) and textures (e.g. continuous length of voxels with similar intensities). [8] These features can be sensitive to differences in image parameters such as slice thickness. [9] Post-reconstruction harmonisation methods have been proposed to adjust for these differences, thus promoting standardised research between centres. [10]

The primary aim of this study was to test the generalizability of a UK single-centre esophageal cancer prognostic model incorporating radiomic features [11] firstly pre-harmonisation, then post-harmonisation, against a cohort of esophageal cancer patients treated exclusively with neo-adjuvant chemoradiotherapy (NACRT) according to the Dutch NACRT plus surgery versus surgery alone for oesophageal/junctional cancer (CROSS) trial regimen. [12] A widely generalizable prognostic model incorporating radiomic features of primary tumours might offer clinicians complimentary data beyond traditional prognostic factors that will assist treatment decision making and risk stratification. [11, 13] The secondary aim was to compare prognostic models with and without PET image features between cohorts to provide further validation.

Materials & Methods

This study was designed as a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 external independent validation study. [14] A previously published prognostic model had been developed and internally validated in patients with esophageal cancer. Details of model development have been provided in Foley et al. [11] Briefly, the prognostic model had only been evaluated by same-centre

internal validation in patients managed by the South-East Wales Regional Upper Gastrointestinal (GI) Cancer Multi-Disciplinary Team (MDT), United Kingdom. A suitable independent cohort was not accessible at the time of publication. Institutional board review (IRB) approval was granted for the development of the prognostic model (REF 14/WA/1208). The prognostic model was developed as part of a larger study investigating the prognostic significance of image texture analysis in gastro-oesophageal cancer (STAGE), and from here-on will be known as the STAGE cohort. The external validation cohort comprised patients treated with the CROSS regimen in The Netherlands. IRB permission was obtained for the external validation cohort.

Patient cohorts

In total, 449 patients were included in the development and validation of this prognostic model. Figure 1 details the number of patients in each cohort and the reasons for exclusion of patients from the CROSS validation cohort. The largest number of patient exclusions (n=23) from the CROSS cohort were because of the pre-defined metabolic tumour volumes (MTV) adopted in Foley et al [11] and used in this current study for consistency. A sensitivity analysis of these excluded cases has been included in Appendix B. Other main reasons for patient exclusion were different calibration units (n=11) and ATLAAS segmentation failure (n=7).

Primary Outcome

The primary endpoint of the published prognostic model is overall survival, defined as the number of months survived after the date of diagnosis until death or last day of follow-up. Dates of death were obtained from the Cancer National Information System Cymru (CaNISC) database (Velindre NHS Trust, Wales), reported by the Office for National Statistics. Dates of death of patients in the CROSS cohort were obtained from the national registry. In both cohorts, local researchers were not blinded to the dates of death. A uniform and standardised procedure for autosegmentation and radiomics computation was implemented at each centre to ensure consistent methodology.

Tumour Segmentation

Primary tumours were segmented on PET images using an automatic tree-based learning algorithm for advanced segmentation (ATLAAS). [15] The benefit of ATLAAS is that inter-observer variability in contouring is eliminated. Full details regarding the use of ATLAAS in this study are provided in Foley et al. and Berthon et al. [11, 15]

The following model equation (Eq. 1) was used to calculate a prognostic score for each patient. This equation was derived using published methods. [16]

$$\textit{Prognostic score} = \textit{Stage Group}*0.397 - \textit{Treatment}*1.094 + \textit{Age}*0.024 - \log(\textit{Histogram Energy})*1.320 + \log(\textit{TLG})*1.748 + \textit{Histogram Kurtosis}*0.198 \quad \textit{Eq. 1}$$

External Validation

The ATLAAS code and equations to calculate each of the PET image features were shared between institutions. The primary tumours on the PET scans of the CROSS patients were then segmented using ATLAAS and the MTVs produced were visually assessed for adequacy for quality control. Validation was firstly performed with pre-harmonisation metrics and then repeated with post-harmonisation PET features to adjust for potential differences between scanners. Fully anonymised data was then shared between institutions.

Different PET/CT scanners and protocols were used across the cohorts (Appendix A). Radiomics features are known to change significantly as a function of scanner model, image acquisition or reconstruction settings, therefore we explored using the post-reconstruction Combat harmonisation method [17] to harmonise features extracted from images acquired across different scanners. Slice thickness was chosen for harmonisation because images from one scanner had different thickness values, which resulted in 5 categories (Appendix A, Table A.1). Further details of the cohorts, treatments received, PET/CT protocols, metric equations, variation in image features and the post-reconstruction PET harmonisation Combat method [17], used to adjust for batch effects across different datasets, have been provided in Appendix A.

Statistical analysis

Categorical data are described as frequency (percent) and continuous variables as median (range) and differences assessed with appropriate non-parametric tests. There was no missing data in the development cohort and cases with missing data were excluded from the validation CROSS cohort. Patient characteristics at staging were compared for each cohort. Boxplots were generated locally on each cohort to compare the distributions of the model variables. Firstly, the published model was applied to 46 suitable patients in the CROSS cohort prior to PET harmonisation. A second model validation was then performed using image features calculated post-harmonisation. Model discrimination was evaluated using the log-rank test; a p-value of <0.05 was defined as statistically significant. Model calibration followed a standard test procedure detailed in [18], and which has been previously implemented in [19]. In this study, we define model discrimination as preserved if the p-value of the calibration slope $\beta = 1$ is >0.05. Thirdly, we performed the same validation steps for a prognostic model developed on the same STAGE cohort, but exclusively using clinical variables (age at diagnosis, stage and treatment) and no imaging based variables. Statistical analysis was performed with SPSS version 23.0 (IBM, Chicago, USA) and MATLAB version 9.0 (MathWorks, Natick, MA).

Results

The baseline characteristics of the STAGE development, validation and CROSS cohorts are detailed in Table 1. The median overall survival of the CROSS cohort was 25 months (95% confidence interval (CI) 23.0 to 31.4). The median overall survival of the STAGE development and validation cohorts was 16.0 months (95% CI 13.8-18.2) and 14.0 months (95% CI 10.4-17.6), respectively.

Boxplots were constructed to compare the values of log(TLG), log(Histogram Energy) and Histogram Kurtosis in between the STAGE and CROSS cohorts. (Fig. 2) Additional boxplots

and descriptive statistics of PET feature values pre- and post-harmonisation are included in Appendix B. There were similar mean values and distributions of the 3 variables between STAGE and CROSS cohorts, although a greater number of outliers were observed for Histogram Kurtosis in the STAGE cohort. This is probably due to a larger number of patients and greater range in MTV of the primary tumours included in the STAGE cohort. (Table B.1)

A prognostic model containing clinical variables only was calculated from the STAGE development cohort using identical data from the original study. Age at diagnosis (HR 1.025, 95% CI 1.011-1.040, p<0.001), stage (0.337, 0.243-0.468, p<0.001) and treatment (1.462, 1.187-1.802, p<0.001) were all independently and significantly associated with overall survival.

Prognostic model developed by clinical and radiomic features

Pre-harmonisation

Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.27, df=3, p=0.74). (Fig 3) The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.89 (95% CI 0.29-2.75), 1.36 (95% CI 0.47-3.92) and 0.78 (95% CI 0.25-2.41), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 1.09 (standard error (SE) 0.41). β is not significantly different from 1 (p=0.84), which indicates that model discrimination is preserved.

The mean overall survival for patient quartiles 1-4 were 34.0 months (95% CI 19.0-49.2), 29.5 months (95% CI 19.5-39.5), 25.9 months (95% CI 14.8-37.0) and 41.2 months (95% CI 25.9-56.4), respectively. Median overall survival could not be calculated for all quartiles. The median prognostic score for quartiles 1-4 was -0.51 (n=11, range -1.14 to -0.37), -0.15 (n=11, range -0.36 to 0.01), 0.20 (n=11, range 0.04 to 0.30) and 0.48 (n=13, range 0.30 to 1.16), respectively.

Post-harmonisation

Following post-reconstruction PET harmonisation, repeated Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.42, df=3, p=0.70). (Fig 3) The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.78 (95% CI 0.24-2.55), 1.47 (95% CI 0.50-4.25) and 1.15 (95% CI 0.39-3.40), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 1.26 (standard error (SE) 0.22). β is not significantly different from 1 (p=0.29), which indicates that model discrimination is preserved. The adjusted survival data for the patient quartiles is available in Appendix B.

These results indicate that PET harmonisation did not have a substantial effect on model validation, with similar results obtained using both methods.

Prognostic model developed with clinical features only

The median prognostic score of the model developed with clinical variables only was -2.68 (range -4.89 to -0.17). As shown in Figure 4, Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.19, df=3, p=0.75). The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.93 (95% CI 0.27-3.23), 1.41 (95% CI 0.45-4.43) and 1.53 (95% CI 0.51-4.57), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 2.15 (SE 0.72). β is not significantly different from 1 (p=0.29), which indicates that model discrimination is preserved.

In the prognostic model with clinical variables only, patients in quartiles 2 & 3 were combined to create an intermediate risk group, following a previously published method. [20] (Fig. 5) Applying Bonferroni correction, there was no statistically significance difference between the low, intermediate and high risk groups (X^2 0.712, df 2, p=0.701) but a separation in overall survival curves was observed (intermediate risk vs low risk HR 1.16 (95% CI 0.41-3.30 and high risk vs low risk HR 1.53 (95% CI 0.51-4.58)). The calibration slope β= 2.15 (SE .72, p-value 0.29) indicating model discrimination was preserved.

Discussion

Patients with esophageal cancer have a poor prognosis and the incidence of the disease is increasing. [21] Despite advances in modern healthcare, survival rates remain low. Enhanced staging algorithms are required to improve the accuracy of staging, which informs clinicians of the likely prognosis and provides subsequent patient risk stratification. Prognostic models incorporating radiomic features are one strategy being investigated for this purpose.

This external validation study has shown that results of a developed prognostic model combining clinical risk factors and PET radiomics features was not replicated in a cohort of patients treated with the CROSS trial regimen. However, when a prognostic model including only clinical variables from the STAGE development cohort was tested, some aspects of the model were indicative of transferability to the CROSS cohort. Our data shows that clinical features of esophageal cancer remain prognostic across different countries and studies.

Despite not being able to replicate the validation results of the published prognostic model, this study remains clinically important because more accurate staging of esophageal cancer is essential to improve survival rates. Validated prognostic and predictive radiomics models are one strategy to improve radiological staging of esophageal cancer. [22] Greater staging accuracy will improve patient risk stratification, which is critically important for optimising personalised treatment decision-making. Once validated, staging algorithms incorporating radiomics may enable clinicians to decide upon the best management plan from the outset of diagnosis, therefore providing the greatest chance of survival for each patient.

A number of important methodological reasons in the modelling process may have contributed to the lack of external validity of the prognostic model when transported to the CROSS observations. First, the PET image acquisition protocols in the CROSS regimen cohort may not have been as strictly policed as in the STAGE study, leading to divergence in PET acquisition parameters. (Table A.1) All patients in STAGE (n=403) were staged using the

same PET/CT scanner and protocol. However, different PET/CT scanners and protocols were used in both the STAGE and CROSS cohorts. Harmonising PET image features demonstrated little improvement in the model validity between cohorts.

Harmonising PET image features demonstrated little improvement in the model validity between cohorts, which supports this post-reconstruction method in external validation radiomics studies and suggests that harmonisation had little influence in these cohorts. These findings contradict those of Orlhac et al. [10] Several factors could explain the lack of effect. The clinical variables of patient age, TNM stage and treatment are likely to have the greatest impact on overall survival compared to the image features. The PET features used in the original model by Foley et al (TLG, Histogram Energy and Histogram Kurtosis) were not investigated in Orlhac et al. Furthermore, although the Combat algorithm has been used in genomics, it has not yet been validated in radiomics. A consensus on uniformly standardised PET imaging protocols is required for multi-institutional validation of prognostic/predictive models incorporating radiomics. [23]

Second, the prognostic model excluded patients with small MTV < 5 mL, thus further reducing the number of CROSS patients that were eligible for validation. The small patient numbers in the external validation cohort limits the ability to replicate the results of the STAGE prognostic model. This study is likely to be under-powered and improved validation could be achieved by increasing the cohort size. Patients with a smaller MTV were more likely to be suitable for radical therapy and therefore eligible for recruitment into the CROSS trial. When the excluded small MTV cases were tested in the sensitivity analysis included in Appendix B, no significant difference in overall survival between patient quartiles remained (X^2=3.85, df=3, p=0.28). In addition, evidence at the time of prognostic model development suggested possible unstable segmentation at smaller MTVs and an increase in redundant (highly cross-correlated) radiomic data that can be extracted. [24] There is no clear consensus on minimum MTV in PET radiomics studies. One study recommends excluding MTVs of < 45 mL, although only one calculation choice for local entropy, despite the many possibilities of discretisation steps and matrices available, was evaluated in this study. [25] Other studies have previously recommend excluding patients with a primary MTV of < 10 mL. [26, 27] However, prognostic models including image features extracted from small tumour volumes can still be developed. [8] The original model by Foley et al. did not examine a wide range of higher order features, some of which may have turned out reproducible and significantly prognostic with the expanded dataset. However, since the scope of this study was only the feasible generalizability of the original model, we did not re-analyse using additional textural features. The possibility for including redundant data exists but providing the study is appropriately powered, the model can still be compared to those containing only clinical variables.

Third, the development of the previous prognostic model did not include an exhaustive radiomic feature selection steps to identify features that would be robustly reproducible within the STAGE cohort and hence more likely to be transferable to the CROSS cohort. [8] Details of the PET variables implemented in the developed prognostic model can be found in Foley et al. [11] These variables were shown to have prognostic significance in the early radiomics literature [28-30] and were implemented identically.

More studies are required to test the reliability, robustness and additional value of PET image features across a range of MTVs and between different PET/CT scanners. [9, 26] Regarding the original model, TLG and Histogram Energy have shown good reproducibility results, however there is mixed evidence for Histogram Kurtosis. [31] Previous studies have found significant associations between higher order features and overall survival [29] and that the amount of complementary radiomic information gained increases with larger MTVs. [26] Despite this, the original development study did not demonstrate prognostic significance of any higher order features, although only 3 such features were investigated.

Advanced correction algorithms are being developed to harmonise features extracted from scans with different acquisition parameters, which could greatly benefit multi-centre radiomic studies and reduce variation in metrics. [32]

Standardisation efforts such as the Image Biomarker Standardisation Initiative (IBSI) [33] are an important methodological step towards reducing sensitivity of radiomic features to computation (image extraction) software. Deployment of the same autosegmentation tool (ATLAAS [15]) reduced inter-observer variability in contouring and the same feature extraction software that was executed locally was used in both participating centres. These techniques are examples of standardised processes that improve the robustness of radiomic features.

Lastly, a relatively small proportion of the STAGE cohort received NACRT or surgery alone (Table 1). These differences may not have been adjusted for completely by the original model multivariate regression. The STAGE cohort is relatively heterogeneous cohort of patients compared to the CROSS cohort, because it was collected during an observational cohort study recruiting all patients with esophageal cancer. Patients in the CROSS cohort were all treated with NACRT, so they share more similar characteristics. Differences between validation cohorts are important in external validation studies because the generalisation of the model can be tested at its extremes. Furthermore, this points the way forward to improved (reproducible) feature selection methodology and updating of the original model to address a more generalized clinical question.

All prognostic models must be validated in an independent external cohort before being considered for use in clinical practice because many models present optimistic and over-fitted results from development cohorts. [34] However, external validation studies are rarely performed. A review of the performance of prognostic models showed that 11% are externally validated. [35] This may explain why few developed prognostic models are adopted into clinical practice. [36] Our collaborative research group is planning to update this prognostic model and perform a further external validation study with more robust feature selection and standardised feature extraction algorithms using all tumour volumes.

In conclusion, this initial TRIPOD type 3 external validation study evaluated a prognostic model developed in esophageal cancer patients staged with PET/CT. The prognostic model did not enable significant discrimination between patient risk groups in the CROSS cohort, but a second model including clinical variables only (age, disease stage and treatment) demonstrated transferable prognostic factors between international cohorts.

Acknowledgements
The authors wish to acknowledge the contributions of Professor Robert K Hills who developed the original prognostic model, Professor Wyn G Lewis who helped with the STAGE cohort data collection, Professor Christopher Marshall (Director of the Positron-Emission Tomography Imaging Centre (PETIC) in Cardiff and members of the South-East Wales Upper GI Cancer MDT committee.

Ethical Statement
Institutional review board approval was obtained.

Data Availability
The data that has been used in this study is confidential and cannot be shared

Funding
The study was partially funding by a UK Tenovus Cancer Care Grant (TIG2016/04).

Competing interests
The authors declare that they have no competing interests.

Author contributions
KF, AR, LW and ES conceived and designed the study. RL, MB, MS, PK and TC collected the data. ZS, PW, CP and PK preformed the data analysis. KF, LW, JS, TC and AD drafted the manuscript. All authors read and approved the final manuscript.

References

[1] Cancer Research UK. Oesophageal Cancer Statistics. 2016 [Accessed November 22nd 2016]; Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer.
[2] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-86.
[3] Kayani B, Zacharakis E, Ahmed K, Hanna GB. Lymph node metastases and prognosis in oesophageal carcinoma-a systematic review. Eur J Surg Oncol. 2011;37:747-53.
[4] Foley KG, Christian A, Fielding P, Lewis WG, Roberts SA. Accuracy of contemporary oesophageal cancer lymph node staging with radiological-pathological correlation. Clin Radiol. 2017;72:e691-e7.
[5] McCormick Matthews LH, Noble F, Tod J, Jaynes E, Harris S, Primrose JN, et al. Systematic review and meta-analysis of immunohistochemical prognostic biomarkers in resected oesophageal adenocarcinoma. Br J Cancer. 2015;113:107-18.
[6] Cook GJR, Siddique M, Taylor BP, Yip C, Chicklore S, Goh V. Radiomics in PET: principles and applications. Clin Transl Imaging. 2014;2:269-76.
[7] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441-6.
[8] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
[9] Desseroit MC, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort. J Nucl Med. 2017;58:406-11.
[10] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018.
[11] 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. Eur Radiol. 2018;28:428-36.
[12] van Hagen P, Hulshof MCCM, van Lanschot JJB, Steyerberg EW, van Berge Henegouwen MI, Wijnhoven BPL, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. 2012;366:2074-84.
[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. Eur Radiol. 2018.
[14] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
[15] Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. Phys Med Biol. 2016;61:4855-69.

[16] Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683-90.
[17] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118-27.
[18] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
[19] Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol. 2015;54:1423-9.
[20] Dekker A, Vinod S, Holloway L, Oberije C, George A, Goozee G, et al. Rapid learning in practice: a lung cancer survival decision support system in routine patient care data. Radiother Oncol. 2014;113:47-53.
[21] Cancer Research UK. Oesophageal Cancer Incidence Statistics. 2016 [Accessed December 20th 2016]; Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer/incidence.
[22] van Rossum PS, Xu C, Fried DV, Goense L, Court LE, Lin SH. The emerging field of radiomics in esophageal cancer: current evidence and future potential. Transl Cancer Res. 2016;5:410-23.
[23] 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. Med Phys. 2017;44:e1-e42.
[24] Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. Front Oncol. 2016;6:71.
[25] Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med. 2014;55:37-42.
[26] Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, et al. 18F-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.
[27] Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med. 2014;55:414-22.
[28] Hatt M, Visvikis D, Albarghach NM, Tixier F, Pradier O, Cheze-le Rest C. Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. Eur J Nucl Med Mol Imaging. 2011;38:1191-202.
[29] Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med. 2011;52:369-78.
[30] Yip C, Landau D, Kozarski R, Ganeshan B, Thomas R, Michaelidou A, et al. Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy. Radiology. 2014;270:141-8.
[31] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys. 2018;102:1143-58.

[32] Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One. 2017;12:e0178524.
[33] Zwanenburg A, Leger S, Vallieres M, Lock S. Image biomarker standardisation initiative - feature definitions. 2016 [Accessed March 20th 2017]; Available from: https://arxiv.org/abs/1612.07003v3.
[34] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19:453-73.
[35] Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Med. 2010;8:21.
[36] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201-9.

Tables

Table 1. Baseline Characteristics of Patients in Development, Validation and CROSS Cohorts

Frequency (%)	**STAGE Development Cohort (n=302)**	**STAGE Validation Cohort (n=101)**	**CROSS cohort (n= 46)**	**p-value***
Median Age	67.0 years (Range 39-83)	69.0 years (Range 39-84)	64.5 years (Range 47-77.8)	0.114
Gender (M: F)	227 (75.2): 75 (24.8)	78 (77.2): 23 (22.8)	38 (82.6): 8 (17.4)	0.528
Histology				0.602
Adeno	237 (78.5)	79 (78.2)	39 (84.8)	
SCC	65 (21.5)	22 (21.8)	7 (15.2)	
Tumour Location				0.010
Oesophagus	192 (63.6)	47 (46.5)	28 (60.9)	
Gastro-oesophageal junction	110 (36.4)	54 (53.5)	18 (39.1)	
Stage Groups				0.018
Stage 1	17 (5.6)	2 (2.0)	2 (4.4)	
Stage 2	56 (18.5)	24 (23.8)	10 (21.7)	
Stage 3	160 (53.1)	57 (56.4)	33 (71.7)	
Stage 4	69 (22.8)	18 (17.8)	1 (2.2)	
Treatment				<0.001
Curative	158 (52.3)	50 (49.5)	46 (100)	
SA	24 (15.2)	4 (8.0)	0 (0.0)	
NACT	67 (42.4)	23 (46.0)	0 (0.0)	
NACRT	13 (8.2)	7 (14.0)	46 (100)	
dCRT	54 (34.2)	16 (32.0)	0 (0.0)	
Palliative	144 (47.7)	51 (50.5)	0 (0.0)	
Overall Survival				<0.001
Alive	70 (23.2)	43 (42.6)	20 (43.5%)	
Dead	232 (76.8)	58 (57.4)	26 (51.5%)	

SCC squamous cell carcinoma; SA surgery alone; NACT neo-adjuvant chemotherapy; NACRT neo-adjuvant chemoradiotherapy; dCRT definitive chemoradiotherapy; *chi-square test

Figure Legends

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

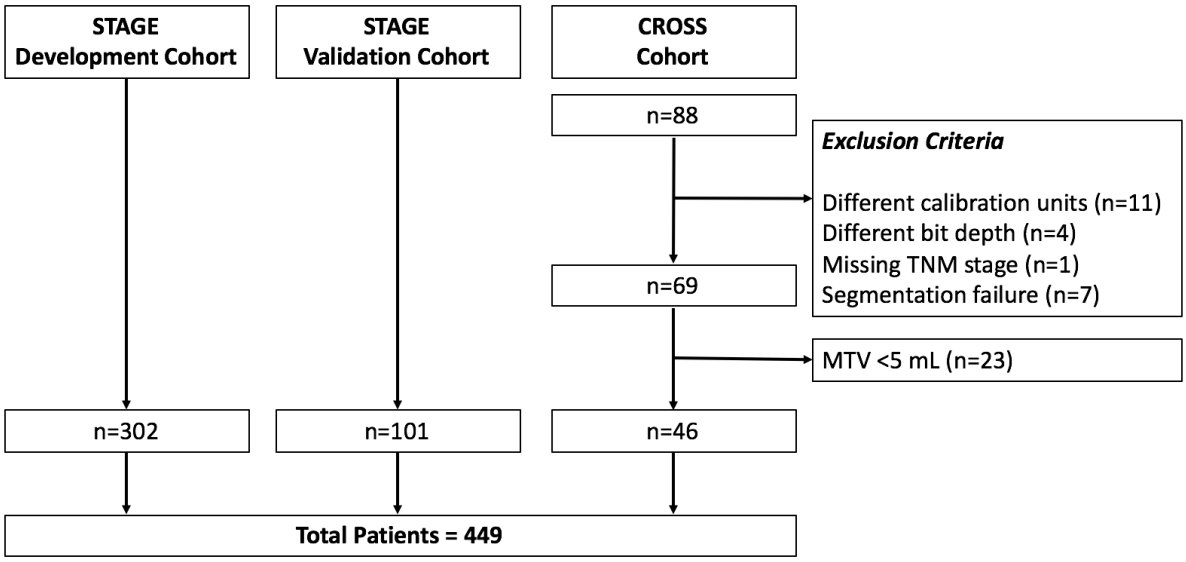


Figure 2. Boxplots displaying pre-harmonisation mean values and interquartile ranges of log(TLG), log(Histogram Energy) and Histogram Kurtosis in STAGE and CROSS cohorts.

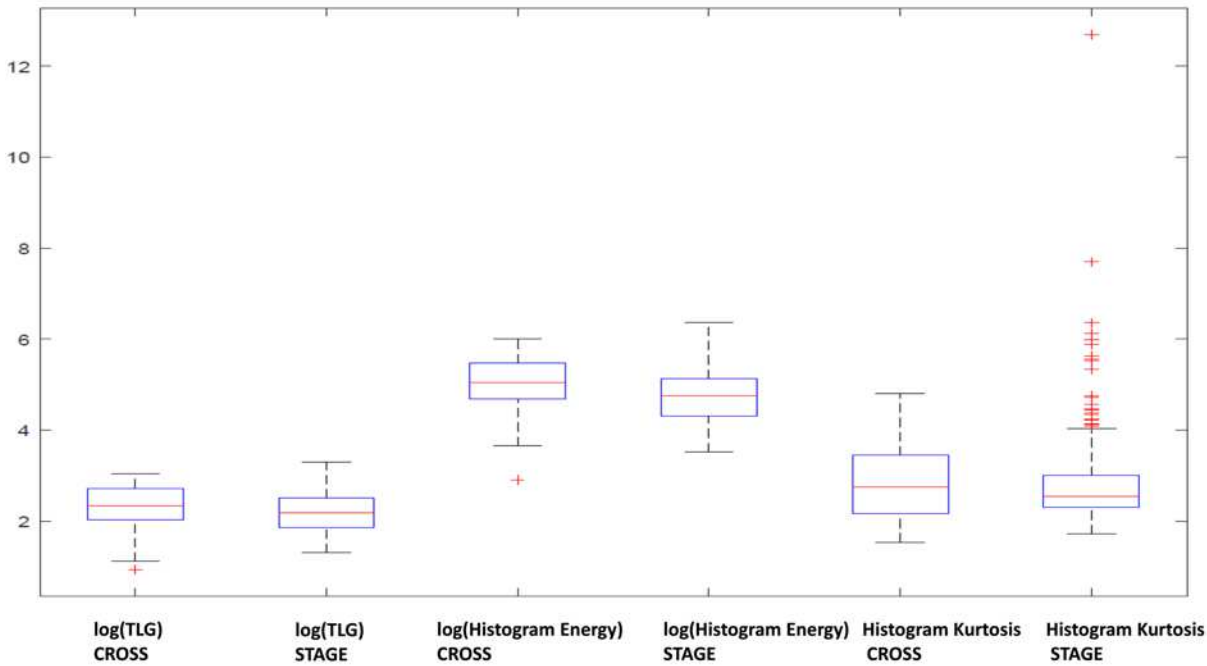
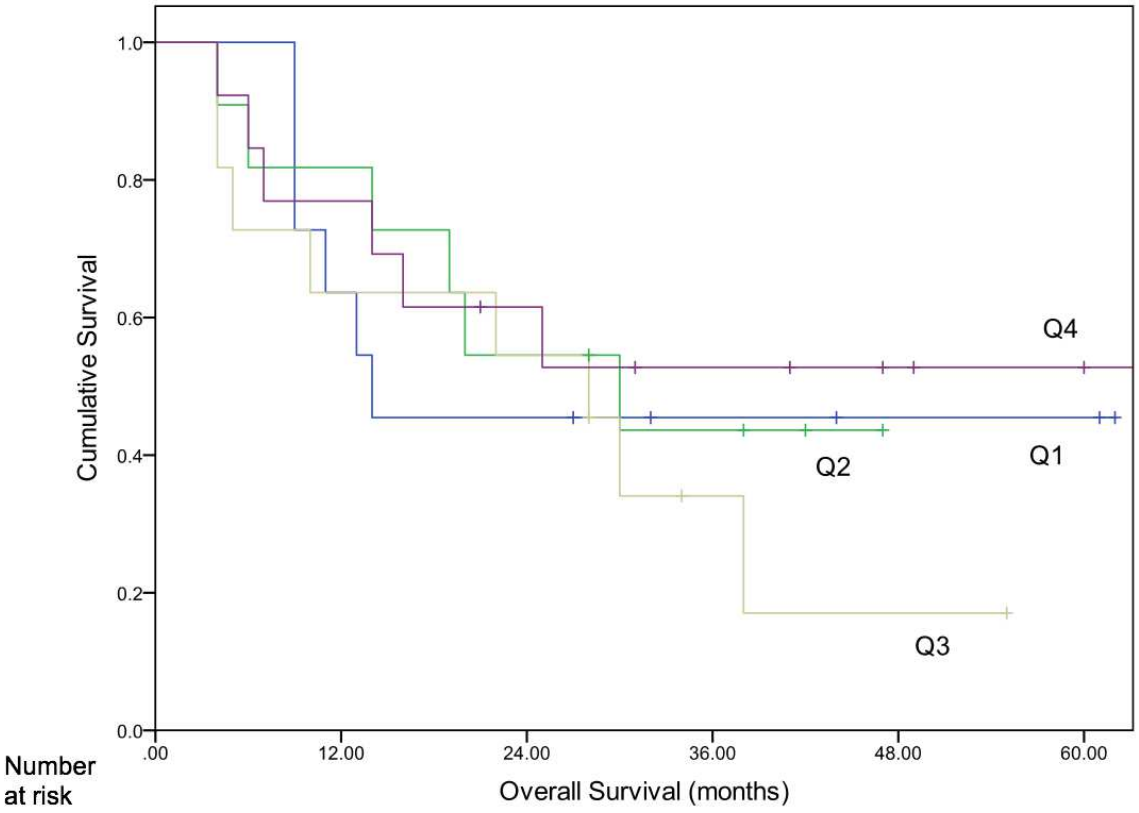
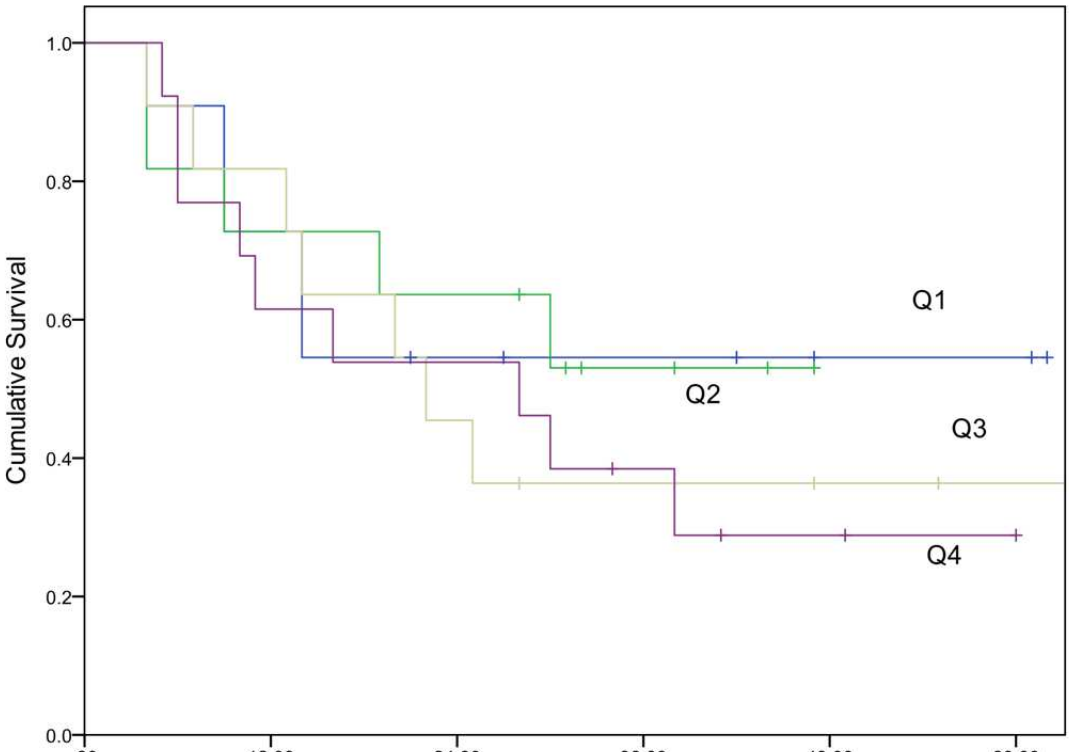


Figure 3. Cumulative survival curves of patient quartiles (Q1-4) in CROSS cohort using model developed with clinical and radiomic features (X^2=1.27, df=3, p=0.74).



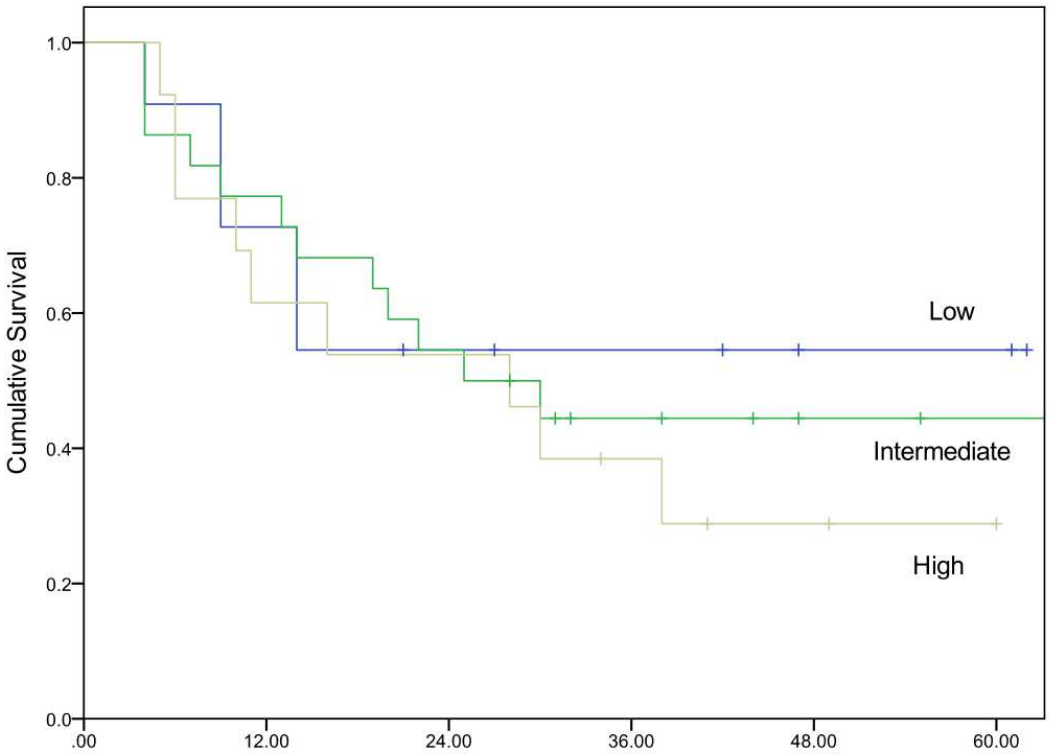
Number at risk						
Q1	11	7	5	3	2	2
Q2	11	9	6	4	0	0
Q3	11	7	6	2	1	0
Q4	13	10	7	5	3	1

Figure 4. Cumulative survival curves of patient quartiles (Q1-4) in CROSS cohort using model developed with clinical features only (X^2=1.19, df=3, p=0.75).



Number at risk						
Q1	11	8	6	6	6	6
Q2	11	8	7	6	6	6
Q3	11	9	5	4	4	4
Q4	13	8	7	5	4	4

Figure 5. Cumulative survival curves of combined risk groups in CROSS cohort using model developed with clinical features only. The original quartile 1 corresponds to the low-risk group, quartiles 2 & 3 were combined to create an intermediate risk group and quartile 4 corresponds to the high-risk group.



Number at risk						
Low	10	8	5	3	2	2
Intermediate	21	17	12	6	2	1
High	12	8	7	4	2	0