Radiotherapy and Oncology 193 (2024) 110084

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy

Joshua R. Astley [a], James M. Reilly [a], Stephen Robinson [a], Jim M. Wild [a,b], Matthew Q. Hatton [a], Bilal A. Tahir [a,b,*]

[a] *Division of Clinical Medicine, The University of Sheffield, Sheffield, UK*
[b] *Insigneo Institute for in Silico Medicine, The University of Sheffield, Sheffield, UK*

ARTICLE INFO

Keywords:
Explainable deep learning
Non-small cell lung cancer
Radiotherapy
Survival analysis

ABSTRACT

Background and purpose: Survival is frequently assessed using Cox proportional hazards (CPH) regression; however, CPH may be too simplistic as it assumes a linear relationship between covariables and the outcome. Alternative, non-linear machine learning (ML)-based approaches, such as random survival forests (RSFs) and, more recently, deep learning (DL) have been proposed; however, these techniques are largely black-box in nature, limiting explainability. We compared CPH, RSF and DL to predict overall survival (OS) of non-small cell lung cancer (NSCLC) patients receiving radiotherapy using pre-treatment covariables. We employed explainable techniques to provide insights into the contribution of each covariable on OS prediction.
Materials and methods: The dataset contained 471 stage I-IV NSCLC patients treated with radiotherapy. We built CPH, RSF and DL OS prediction models using several baseline covariable combinations. 10-fold Monte-Carlo cross-validation was employed with a split of 70%:10%:20% for training, validation and testing, respectively. We primarily evaluated performance using the concordance index (C-index) and integrated Brier score (IBS). Local interpretable model-agnostic explanation (LIME) values, adapted for use in survival analysis, were computed for each model.
Results: The DL method exhibited a significantly improved C-index of 0.670 compared to the CPH and a significantly improved IBS of 0.121 compared to the CPH and RSF approaches. LIME values suggested that, for the DL method, the three most important covariables in OS prediction were stage, administration of chemotherapy and oesophageal mean radiation dose.
Conclusion: We show that, using pre-treatment covariables, a DL approach demonstrates superior performance over CPH and RSF for OS prediction and use explainable techniques to provide transparency and interpretability.

Introduction

Although the use of radiotherapy (RT) as part of a multimodality treatment approach has improved outcomes for patients with non-small cell lung cancer (NSCLC), survival rates continue to lag behind those seen for many other cancers. The Radiation Therapy Oncology Group 0617 observed that a high-dose regime of 74 Gray (Gy) in 37 fractions surprisingly exhibited worse overall survival (OS) when compared to the standard-dose regime of 60 Gy in 30 fractions in NSCLC patients [1]. The decreased OS has been attributed to increased radiation dose to the heart; incidental radiation doses to other organs-at-risk (OAR) such as the lungs or oesophagus, which can result in radiation-induced pathologies such as pneumonitis or oesophagitis, respectively, have also been associated with decreased OS in NSCLC patients receiving RT [1–3].

OS is frequently assessed using Cox proportional hazards (CPH) regression; however, this method may be too simplistic as it assumes a linear relationship between features (i.e. independent variables) and the outcome (i.e. the dependent variable) [4]. Machine learning (ML) methods have been adapted for use in survival analysis, including random survival forests (RSFs) and gradient boosting methods [5,6]; these approaches can capture higher-order representations compared to CPH. Recently, deep learning (DL) has been proposed as an alternative method to capture more complex, non-linear associations between features and the outcome [7]. DeepSurv is a CPH-based feed-forward neural network [8] which has shown improved performance over CPH and ML techniques in various applications [9–11]. These survival analysis

* Corresponding author at: 18 Claremont Crescent, Sheffield S10 2TA, UK.
E-mail address: b.tahir@sheffield.ac.uk (B.A. Tahir).

https://doi.org/10.1016/j.radonc.2024.110084
Received 3 August 2023; Received in revised form 20 December 2023; Accepted 2 January 2024
Available online 18 January 2024

methods aim to model the survival function $S(t)$ which denotes the probability that a patient is either dead or alive by a given time t and is calculated as:

$$S(t) = exp(-H(t)) \tag{1}$$

where $H(t)$ represents the cumulative distribution, or cumulative hazard, function (CHF) and is related to the probability density function, or hazard, function, $h(t)$ as follows:

$$H(t) = \int_0^t h(x)dx \tag{2}$$

Despite the success of various survival prediction models, these models can largely be considered as 'black boxes' i.e. they are uninterpretable with respect to the importance of each feature on the output prediction. This is undesirable as the prediction is not well understood, leading to a lack of trust by clinicians and patients alike.

Several methods have been proposed to provide global and local explanations for ML and DL models; these include local interpretable model-agnostic explanations (LIME) values and Shapley additive explanations (SHAP) values which assess the local and global importance of features, respectively [12,13]. LIME values derive explanations locally by perturbing the dataset using synthetic feature values randomly generated in the neighbourhood of the specific testing case. LIME values are agnostic to the specific model that is to be explained; only the model's input in the form of a set of features and its output is required to generate model explanations. In contrast, SHAP values derive explanations globally using a game-theoretic framework which interprets each feature as a player in a game where a feature's individual contribution to the final 'pay-out' or prediction is assessed. Therefore, SHAP values are computed using the whole training set and not individual testing set examples, requiring the trained model to provide explanations. Both LIME and SHAP values have been adapted for use in survival analysis [14,15].

In this study, we compared the conventional CPH model with the ML-based RSF and the DL-based DeepSurv models to predict OS from baseline features in NSCLC patients receiving radical RT. We assessed the effects of a plethora of feature combinations on OS prediction. Additionally, we used explainable frameworks to interpret the importance of covariables for each survival model.

Materials and methods

Study participants and dataset generation

The dataset contained clinical, demographic, treatment, and time-to-event survival data for 471 NSCLC patients treated with radical RT between January 2010 and October 2016. Patient data used within this manuscript received ethical approval from the relevant institutional review board. Appropriate consent and permissions have been granted by the sponsors to utilise this data for retrospective purposes. All data were pseudo-anonymised, and all investigations were conducted in accordance with the appropriate guidelines and regulations. Covariables were collected for each patient using medical records from Western Park Cancer Centre, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK; these covariables included age, sex, stage (TNM version 7), administration of chemotherapy, neutrophil–lymphocyte ratio (NLR), planning target volume (PTV) and the mean radiation dose to the heart, lungs, oesophagus and spinal cord. In addition to these covariables, the date of death was recorded. The date of censor for the patients still alive was the date of their most recent follow-up in medical records as of September 2021.

Patients were treated with hypofractionated accelerated RT or continuous hyperfractionated accelerated RT (CHART). Hypofractionated accelerated RT involved 55 Gy in 20 fractions over four weeks. CHART involved 54 Gy in 36 fractions over 12 days. We have previously published work, using an overlapping dataset to the one used in this study, which indicates that there are no significant differences in OS between CHART and hypofractionated accelerated RT [16]. The dose per fraction was not adjusted in relation to the RT regime used. When using the standard biological equivalent dose (BED) calculation, CHART is expected to have a lower tumour dose and organ-at-risk dose compared to the hypofractionated accelerated RT regime. However, this does not match with our published clinical experience of equivalent disease-free survival and overall survival with minimal differences in the rate of toxicities between these regimes, including rates of all toxicities and rates of grade 3+ toxicities [16]. Using a 'daily received dose' of 4.5 Gy in 12 fractions provides updated BED values of 78 Gy and 135 Gy for tumour and organ at risk doses, respectively; given this broad range of BED which spans greater and lesser BED values than the hypofractionated accelerated RT regime, we have not adjusted our dose for the different dose per fraction of the two regimes. For patients undergoing chemotherapy, regimens were platinum-based doublets with gemcitabine, vinorelbine or pemetrexed.

Organs of interest were delineated using the Varian Eclipse treatment planning software (Varian Medical Systems, Palo Alto, CA, USA) on each axial slice of the planning computed tomography scan. In accordance with trial quality assurance guidelines [17], hearts were manually delineated from the pulmonary artery to the heart apex and oesophagi were manually delineated from the cricoid cartilage to the gastro-oesophageal junction. Lung structures, without the gross tumour volume, were automatically generated using the Varian Eclipse lung segmentation algorithm and subsequently manually corrected. The PTV was generated by adding a uniform margin of 5 mm to the gross tumour volume and then further expanded by 15 mm craniocaudally and 10 mm axially. The mean planned doses to the heart, lungs, oesophagus and spinal cord were retrieved for each patient.

Data split

Monte-Carlo cross-validation with 10-folds was performed using all 471 patients; the dataset was divided using data split percentages of 70:10:20 for training, validation and testing, respectively. This resulted in 338 training, 38 validation and 94 testing patients for each cross-validation fold. Monte-Carlo cross-validation is a type of repeated random sub-sampling where for each data split defined above, training data is used to fit models, parameterisation is conducted to maximise performance on the validation dataset, and the testing set used to validate proposed survival analysis models. By utilising Monte-Carlo cross-validation, the proportion of the training/validation/testing split is not dependent on the number of iterations or partitions; however, some patients may not be part of the testing set at all whereas other participants may be selected multiple times.

Feature selection

In this work, we compared different combinations of input features by varying the covariables included in survival prediction models. Several covariables were selected as features which would be included 'as standard' in all survival models, namely, age, sex, stage, administration of chemotherapy, NLR, PTV and spinal cord mean dose. We then added OAR dosimetric variables, namely, the heart, lung and oesophageal mean dose, resulting in three new feature combinations; these features are added to the 'standard' features. Additionally, the standard features are combined with all OAR dosimetric variables, resulting in a feature combination where 'all' covariables are included.

Survival analysis models

Three survival prediction models were trained. These included the conventional CPH, the ML-based RSF and the DL-based DeepSurv; these frameworks and the specific parameters used are described in detail below.

Some features were pre-processed before entering the survival prediction models. This included the stage of the participant which was ordinally encoded. Additionally, all non-categorical features were standardised to reduce large variations in non-transformed values between features.

Cox proportional hazards

CPH is a semi-parametric approach that models the patient-specific hazard function [4]. CPH assumes that the time component and feature component are proportional with respect to some weighting. As such, the proportional hazard function $h(t|\boldsymbol{X}_i)$ can be defined as:

$$\begin{aligned} h(t|\boldsymbol{X}_i) &= \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_n X_{in}) \\ &= \lambda_0(t)\exp(\boldsymbol{X}_i \cdot \boldsymbol{\beta}_i) \end{aligned} \tag{3}$$

where $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{in})$ is a vector representing the covariables for a patient i with n covariables, $\lambda_0(t)$ denotes the baseline hazard function and $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \beta_{in})$ is a vector that represents the patient-specific coefficients. From Equation (3), it can be inferred that as the baseline hazard function is consistent between patients and is the only time-dependent component of $h(t|\boldsymbol{X}_i)$, the proportional difference between patients is only dependent on the baseline scaling factor $\exp(\boldsymbol{X}_i \cdot \boldsymbol{\beta}_i)$.

CPH models were trained and evaluated in python 3.9 using the scikit-survival sksurv framework with a learning rate of 0.01 and L2 regularisation of 1×10^{-2} with early stopping to limit the number of training epochs.

Random survival forest

RSF is an ensemble model consisting of a series of decision trees adapted to accommodate censored data [5]. RSFs are built by first drawing a set of N random samples from the dataset. A survival tree is grown for each of these N samples creating n trees. At each node of the tree, a subset of covariables x is randomly selected from all features using a split factor y, where y is a specific unique combination of x. The node is split whereby the survival difference between daughter nodes is maximised by searching over all possible x and y values until values are found that maximise the survival difference. This is repeated recursively until a stopping criterion is met, namely, that a terminal node has no less than $d_0 > 0$ unique deaths. Once completed, a CHF is computed for each tree and averaged over n trees to obtain the ensemble CHF. Using the ensemble CHF, the prediction error can be computed using only the data originally first drawn from the overall dataset.

RSF models were trained and evaluated in python 3.9 using the scikit-survival sksurv framework. The number of trees is limited to 200 with a max depth of 5 and minimum node size of 20. Early stopping was employed to limit the number of training epochs.

DeepSurv

DeepSurv is a non-linear implementation of the CPH model utilising a multi-layer feed forward neural network to model non-linear relationships between covariables [8]. This has the potential to produce more accurate survival predictions than the linear relationships modelled by the traditional CPH algorithm. Similar to Equation (3), the hazard function $h(t|\boldsymbol{X}_i)$ can be defined as:

$$h(t|\boldsymbol{X}_i) = \lambda_0(t)\exp(\psi(\boldsymbol{X}_i)) \tag{4}$$

where ψ is a non-linear loss function, determined by the weights of the neural network, and optimised using a gradient descent-based algorithm. The network structure comprised of an input layer, two hidden layers and an output layer with hidden layers of 32 and 16 nodes, respectively. A batch size of 88 was utilised where batch normalisation and a dropout of 0.1 was employed at each layer of the network. A LeakyReLU activation function, negative log-likelihood CPH-based loss function and adam optimisation with a learning rate optimiser was used. Early stopping was employed to limit potential overfitting. DeepSurv models were trained and evaluated in python 3.9 using the pycox framework.

Evaluation metrics

Two primary evaluation metrics were used to assess OS prediction accuracy, namely, the Harrell's concordance index (C-index) and the integrated Brier score (IBS). The C-index is defined between 0 and 1 where 1 corresponds to a perfect prediction and 0.5 corresponds to a random prediction. An IBS of 0.25 corresponds to a 'fence-sitting' prediction and an IBS < 0.25 represents a useful model, whereby the lower the IBS, the more accurate the model prediction. Further details of the evaluation metrics used are given in the Supplementary Material.

Statistical analysis

Statistical analysis was performed using GraphPad Prism 9 (GraphPad, San Diego, CA, USA). A p-value of < 0.05 was considered statistically significant. Univariable and multivariable CPH models were assessed to indicate significant hazard ratios for all features in the dataset; the forced-entry method was used for multivariable analysis. Friedman tests with *post-hoc* multiple comparisons were used to assess significances of differences between different variable combinations for each survival model using both C-index and IBS. Furthermore, Friedman tests with *post-hoc* multiple comparisons were used to assess significances of differences between the best performing variable combinations for each model using both C-index and IBS. To evaluate the methodological risk of bias, we employed the PROBAST tool [18,19]; completed PROBAST forms are available in the Supplementary Material and indicate minimal risk of bias overall.

Results

A summary of covariables alongside hazard ratios and corresponding p-values for both univariable and multivariable CPH analyses are displayed in Table 1. Of the 471 NSCLC patients included, 429 (91%) were deceased and 42 (9%) were recorded as alive at the time of the last follow-up. The median (range) number of survival days, calculated between the date of first RT fraction and last follow-up, was computed as 647 (16, 3794) days.

The training process resulted in 150 individually trained models spread across three survival prediction methods and five covariable combinations. Table 2 indicates the C-index and IBS for the CPH, RSF and DL models. DeepSurv yielded the highest C-index, achieving an average C-index of 0.670 across all testing set cases. Additionally, DeepSurv generated the most accurate IBS, achieving an average IBS of 0.121.

A statistical comparison of each survival model was conducted to assess differences between feature combinations, as displayed in Fig. 1a. Based on these comparisons, the CPH model with 'All' features was determined as the best-performing CPH model. The RSF model with the 'Standard + heart mean dose' feature combination was determined as the best-performing RSF model. For DeepSurv, the configuration with 'All' features was determined as the best-performing DeepSurv model. The best-performing configurations for each approach are displayed in Fig. 1b. Statistical comparisons indicated that using the C-index, DeepSurv significantly outperformed the CPH model; no other significant differences were observed. Using the IBS, DeepSurv significantly outperformed the CPH and RSF models. For the remainder of the Results, these best-performing feature combinations are used in subsequent comparisons.

IBS values are calculated across all time points contained within the testing set; thus a graph of Brier scores, indicating at which time points the Brier score is most accurate, can be generated. Fig. 2 provides a visual representation of the Brier scores over time for the CPH, RSF and DeepSurv approaches. For each approach, the least accurate timeframe

Table 1

Clinical characteristics of the NSCLC patients and univariable/multivariable CPH analyses. Significant variables in the univariable and multivariable CPH analyses are indicated in bold.

Variable		Number (%) or mean ± SD	Univariable Hazard ratio (CI)	Univariable p-value	Multivariable Hazard ratio (CI)	Multivariable p-value
Sex	Male	269 (57)	ref		ref	
	Female	202 (43)	0.837 (0.690, 1.02)	0.071	0.949 (0.773, 1.17)	0.620
Age (years)		71.6 ± 9.19	1.01 (0.995, 1.02)	0.297	**1.01 (1.00, 1.03)**	**0.037**
Stage (%)	I	141 (30)	ref		ref	
	II	68 (14)	**1.40 (1.03, 1.89)**	**0.030**	1.17 (0.844, 1.62)	0.346
	III	234 (50)	**1.37 (1.00, 1.71)**	**0.005**	1.12 (0.795, 1.57)	0.524
	IV	28 (6)	**1.88 (1.24, 2.85)**	**0.003**	1.43 (0.874, 2.35)	0.154
Chemotherapy	No	273 (58)	ref		ref	
	Yes	198 (42)	1.02 (0.843, 1.24)	0.815	**0.734 (0.563, 0.957)**	**0.022**
NLR		3.81 ± 3.36	**1.05 (1.02, 1.08)**	**0.001**	**1.05 (1.02, 1.08)**	**0.002**
PTV (cm^3)		378 ± 219	**1.00 (1.00, 1.00)**	**<0.001**	**1.00 (1.00, 1.00)**	**<0.001**
Heart mean dose (Gy)		10.8 ± 8.19	**1.03 (1.02, 1.05)**	**<0.001**	**1.02 (1.00, 1.04)**	**0.026**
Lung mean dose (Gy)		12.3 ± 4.15	**1.07 (1.04, 1.09)**	**<0.001**	1.04 (0.995, 1.08)	0.087
Oesophageal mean dose (Gy)		14.9 ± 8.72	**1.02 (1.01, 1.03)**	**0.001**	0.983 (0.962, 1.00)	0.110
Spinal cord mean dose (Gy)		7.49 ± 4.61	**1.04 (1.02, 1.06)**	**<0.001**	0.996 (0.966, 1.03)	0.767

Abbreviations: Gy = Gray, CI = confidence interval, SD = standard deviation, cm = centimetre, NLR = neutrophil–lymphocyte ratio, PTV = planning target volume.

Table 2

Performance of CPH, RSF and DeepSurv models in terms of mean (95% CI) C-index and IBS, computed via 10-fold Monte-Carlo cross-validation, for each combination of feature inputs. The best testing set values for each approach are shown in bold; where there was a tie, all tied values are bolded.

Method	Feature combination	CPH		RSF		DeepSurv	
		Training	Testing	Training	Testing	Training	Testing
C-index	Standard	0.586 (0.576, 0.595)	0.601 (0.578, 0.623)	0.658 (0.653, 0.662)	0.645 (0.629, 0.661)	0.664 (0.630, 0.657)	0.630 (0.610, 0.651)
Mean (95 % CI)	Standard + heart mean dose	0.605 (0.593, 0.617)	0.621 (0.599, 0.642)	0.665 (0.662, 0.669)	**0.661 (0.647, 0.674)**	0.638 (0.608, 0.668)	0.628 (0.607, 0.649)
	Standard + lung mean dose	0.600 (0.595, 0.605)	0.619 (0.597, 0.642)	0.666 (0.662, 0.669)	0.649 (0.636, 0.663)	0.637 (0.609, 0.665)	0.628 (0.609, 0.646)
	Standard + oesophageal mean dose	0.588 (0.583, 0.593)	0.607 (0.589, 0.626)	0.662 (0.656, 0.667)	0.645 (0.630, 0.660)	0.646 (0.634, 0.657)	0.642 (0.623, 0.662)
	All	0.629 (0.611, 0.648)	**0.629 (0.611, 0.648)**	0.674 (0.670, 0.678)	0.659 (0.646, 0.672)	0.660 (0.651, 0.668)	**0.670 (0.659, 0.680)**
IBS	Standard	0.134 (0.128, 0.141)	0.145 (0.137, 0.153)	0.110 (0.108, 0.113)	**0.128 (0.120, 0.137)**	0.122 (0.116, 0.128)	0.126 (0.116, 0.137)
Mean (95 % CI)	Standard + heart mean dose	0.135 (0.129, 0.141)	**0.144 (0.135, 0.153)**	0.110 (0.107, 0.112)	**0.128 (0.119, 0.136)**	0.123 (0.115, 0.132)	0.127 (0.116, 0.137)
	Standard + lung mean dose	0.136 (0.130, 0.141)	0.146 (0.136, 0.155)	0.109 (0.107, 0.112)	0.129 (0.120, 0.137)	0.122 (0.114, 0.130)	0.127 (0.116, 0.137)
	Standard + oesophageal mean dose	0.135 (0.128, 0.141)	0.146 (0.138, 0.153)	0.110 (0.108, 0.113)	**0.128 (0.119, 0.137)**	0.119 (0.112, 0.126)	0.123 (0.114, 0.133)
	All	0.139 (0.133, 0.145)	0.146 (0.137, 0.155)	0.108 (0.105, 0.110)	**0.128 (0.119, 0.136)**	0.117 (0.110, 0.124)	**0.121 (0.112, 0.130)**

Abbreviations: C-index = Harrell’s concordance index, IBS = integrated Brier score, CPH = Cox proportional hazards, RSF = random survival forest, CI = confidence interval.

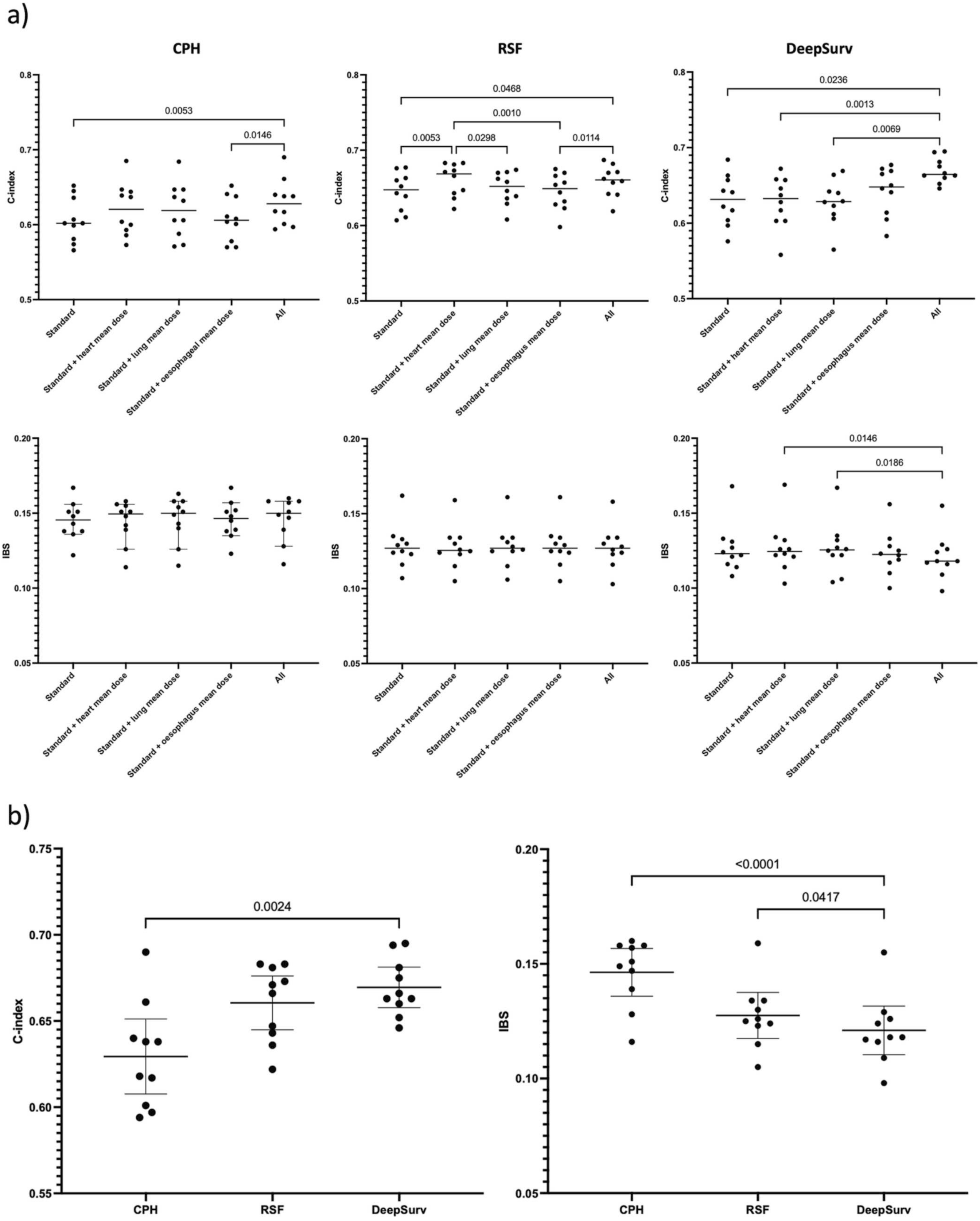

Fig. 1. A) Comparison of trained models with various feature combinations for the cph (left), rsf (middle) and DeepSurv (right) models using the C-index (top) and IBS (bottom). Significant p-values between feature combinations are indicated. b) Test set C-index (left) and IBS (right) values derived from 10 models for the best-performing feature combinations for the CPH, RSF and DeepSurv models. Significant p-values are indicated.

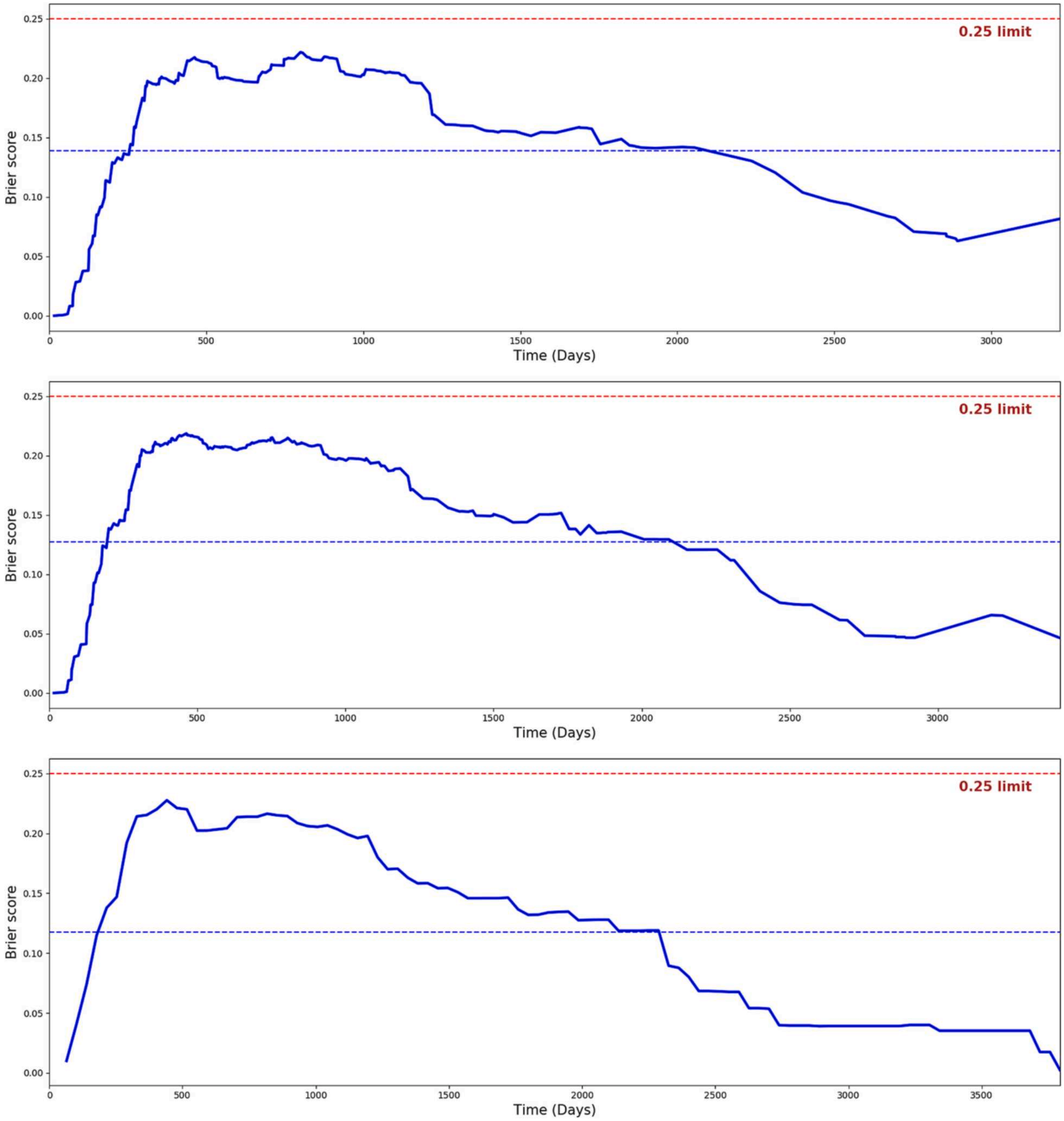

Fig. 2. Brier scores for the CPH (top), RSF (middle) and DeepSurv (bottom) models. The IBS value for each model is indicated by a horizontal blue dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for survival prediction was between 500 and 1500 days.

From Equation (1), the survival function $S(t)$ represents the probability that a patient is alive by time t; predicted survival probability curves generated for each testing set case can be used to calculate the predicted time of death $\widehat{T}$ as follows:

$$\widehat{T} = \int_0^\infty S(t)dt \tag{5}$$

Therefore, the survival error E between the actual survival time T and $\widehat{T}$ is given by $E = \widehat{T} - T$. This error can be used as a quantitative measure to assess the biases of each model, such as over- or under-estimation. Fig. 3 displays the survival error for all testing set cases which experienced an event (i.e. death) for all three survival models tested.

We utilise SurvLIME values to provide feature explanations for survival models used in this work. SurvLIME values are an extension of LIME values that have been adapted for survival analysis [15]. SurvLIME values are generated in python 3.9 using the survlimepy library [20]. Using the 'Monte-Carlo explanation' function, LIME values can be calculated for each testing set case and then the median weighting reported, providing an overall understanding of feature importance for each survival model. Fig. 4 displays median SurvLIME values for the three models utilised in this work for all testing set cases.

RSFs can readily provide feature importance values as nodes are split to maximise the predicted survival difference; therefore, features can be ranked by their corresponding impact at each location they appear, weighted by the number of observations. To further validate SurvLIME values, the RSF's SurvLIME values were compared to its feature

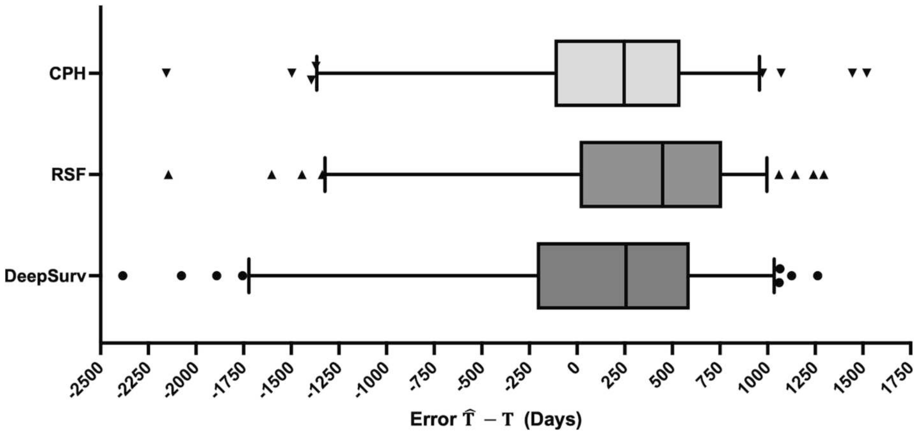

Fig. 3. Differences between actual survival and predicted survival times for testing set cases which had experienced an event (i.e. death) in number of days for the CPH, RSF and DeepSurv models.

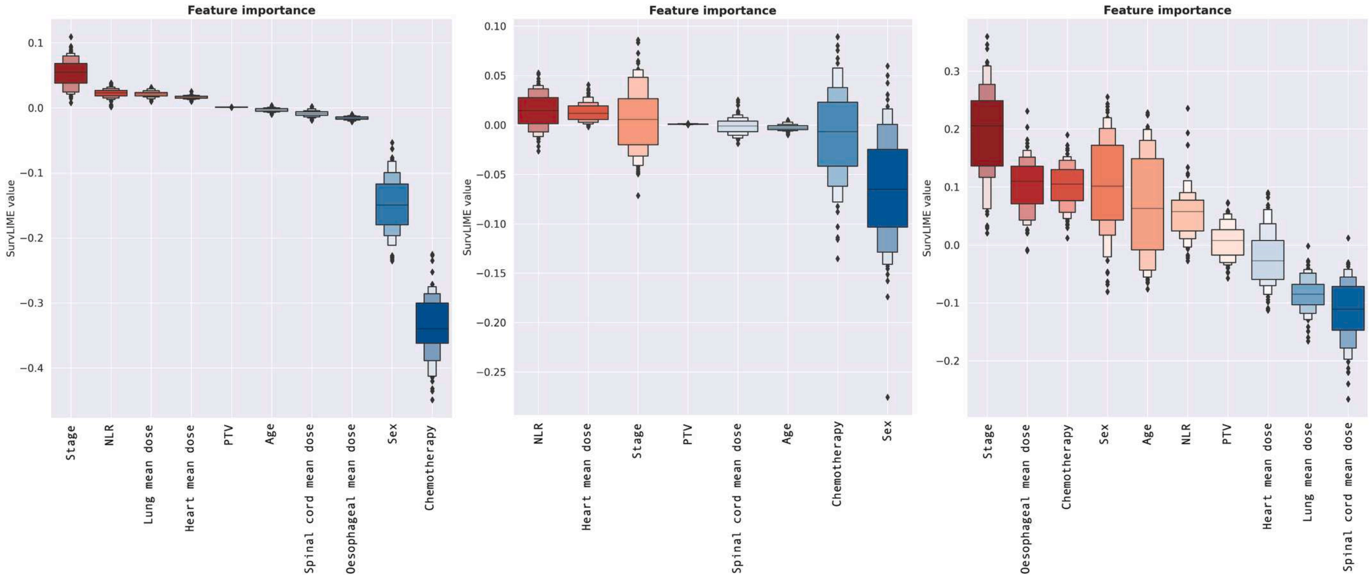

Fig. 4. SurvLIME values for the best-performing CPH (left), RSF (middle) and DeepSurv (right) models.

importance values. A Spearman's correlation of 0.57 was observed, indicating that there is a moderate correlation between both importance measures.

Discussion

Here, we compare several survival models which utilise baseline features to predict OS on a large dataset of NSCLC patients receiving RT, where the DL-based DeepSurv significantly outperformed conventional CPH using the C-index and significantly outperformed CPH and RSF using the IBS. The predicted number of survival days is overestimated by all approaches with the CPH and DeepSurv demonstrating the least survival error.

By computing survival predictions using different feature combinations, the impact of various features can be assessed. For the CPH and RSF models, the oesophageal mean dose appears to exhibit minimal impact on performance, whereas, for DeepSurv, it generated improved performance over other dosimetric features. This potentially indicates that DeepSurv can capture non-linear relationships between the oesophageal mean dose and OS. This is further reinforced by the significant improvement exhibited when all features are utilised, where DeepSurv can capture complex non-linear relationships between inter-related features more effectively than the ML-based RSF. These results indicate that dose-volume features for OARs are influential in accurately predicting OS. Dose to normal tissue in the heart is associated with decreased OS as demonstrated by the RTOG 0617 study [1]. Radiation dose to heart tissue can damage the integrity of endothelial barriers and increases their permeability [21]. For the lungs, radiation dose to lung tissue can lead to radiation-induced lung disease [22]. Additionally, many NSCLC patients have respiratory comorbidities, such as COPD [23]; the additional impact of radiation dose to already damaged tissue can lead to pneumonitis and fibrosis which directly increase the likelihood of clinically significant exacerbations [24]. This has led to the routine use of lung dose constraints in clinical practice to limit the risk of these toxicities for patients. In comparison to other organs, including the heart and lungs, radiation dose to the oesophagus in our centre is generally much less constrained [25]. Consequently, the oesophagus will have greater variation in dose, particularly at the high-dose end of the spectrum, potentially leading to oesophagitis; the frequency of acute oesophagitis has been observed to be increased in accelerated RT regimes [26–28]. As the oesophagus can potentially run very close to the base of the heart, which itself has been shown to be more predictive of OS than the whole heart using CPH [29], it may provide a better reflection of higher dose to that region and, therefore, be particularly useful in survival analysis models.

When comparing the RSF and DeepSurv, using the C-index, we did not observe any significant difference between the two models. Although significant differences are observed using the IBS, the number of features used in the best-performing RSF are reduced when compared to the best-performing DeepSurv feature combination. In particular, the RSF does not require the mean radiation dose to the lungs or oesophagus and, therefore, contouring these OARs, which can be time-consuming, is not required; thus, the marginal benefit in performance exhibited by the DeepSurv must be weighed against clinical throughput. In regard to clinical throughput, it is worth noting that all approaches require the delineation of at least two dose-volume features, potentially limiting the clinical translation of the proposed approaches.

Several researchers have investigated ML- and DL-based models for the prediction of OS in NSCLC patients. Sun *et al.* investigated eight different machine learning models, indicating that the majority of approaches outperformed conventional CPH; the best performing approach, a gradient boosting linear model which used the CPH-based partial log-likelihood method, generated a C-index of 0.68 [30]. In addition, Jin *et al.* studied several ML and DL models to predict OS using a large dataset; however, the dataset is limited to only stage III NSCLC patients [31]. Their proposed approach generated a C-index of 0.83, outperforming both CPH and RSF models. Recently, Lee *et* al. compared several ML- and DL-based survival models for OS prediction in NSCLC patients undergoing RT, utilising a dataset of 428 patients with 29 covariables [32]. They demonstrated that several models improved performance when compared to CPH, achieving a C-index of 0.65, computed over a testing set constituting 49 patients [32]. However, Lee *et* al. utilised some covariables only obtainable after RT (e.g. treatment-related grade 3 $\geq$ adverse events), precluding the model's utility in predicting OS prior to treatment. In comparison, our proposed approach achieved a superior C-index of 0.67 and uses only covariables that can be obtained before RT has commenced; consequently, our approach is deployable earlier in the treatment pathway.

Moreover, Lee *et* al. used explainable techniques, such as SHAP, to interpret model predictions [32]. We employed LIME values to provide insights into the approaches tested, including DeepSurv. LIME values provide local importance information for individual testing set cases; however, in some instances, the local model approximation is dissimilar from a global interpretation of the model in question, reducing the generalisability of explanations. Furthermore, LIME values rely on appropriate perturbations of the dataset to generate accurate predictions; perturbing the dataset is obviously challenging for binary features and, consequently, the prediction of their importance is less accurate than continuous variables. In future work, SHAP values could be calculated alongside LIME values to further validate explanations both from a local and global perspective.

Several limitations are present within the study. The specific covariables utilised are relatively limited; this includes the lack of important predictors, such as performance status, which has demonstrated correlation with OS in NSCLC patients undergoing RT [33,34]; furthermore, other covariables such as smoking status may be important to the prediction of OS. In future work, an expanded number of covariables should be used, alongside increasing the size of the dataset and acquiring data from several centres. Nevertheless, we believe that this work represents a first-step in utilising a more complex, non-linear approach to predicting OS using only baseline variables. A potential use case for the non-linear explainable survival prediction framework we propose could be in the domain of randomised clinical trials, whereby our approach could replace the ubiquitous CPH method.

A potential limitation is the inclusion of spinal cord mean dose as 'standard' whereas the impact of other OAR dosimetric variables were independently investigated. The spinal cord is a purely serial organ and, thus, radiation dose is associated with global organ failure above a certain threshold dose. Radiation myelitis is a late-stage effect that has a profound impact on patient quality of life [35]. To minimise this risk, the use of spinal cord dose constraints are standard practice. Therefore, whilst the spinal cord is considered an OAR, the radiation dose is much more controlled and cannot lead to known adverse conditions which negatively impact survival rates. As shown in Table 1, an increased mean dose to the spinal cord was associated with significantly decreased survival on univariable analysis; however, mean dose to the spinal cord was not significant when mean dose to the heart, lungs and oesophagus were included in multivariable analysis. Despite this, due to the non-linearity of the proposed ML and DL approaches, we believe that, as the forced entry method was used, the mean dose to the spinal cord should still be included.

The selection criteria were kept as broad as possible, encompassing all patients from the original data collection for which all covariables were collected. However, it could be argued that, due to the small number of stage IV patients, differences in plasma osteopontin between stage III and IV, and that stage IV patients receiving radical radiotherapy will be highly selected, that these patients should be excluded from the dataset. We acknowledge these potential limitations; however, we believe that by including all participants the potential use cases of the survival analysis framework is expanded, providing survival predictions which can influence clinical care or palliative treatment. In addition, as all participants were included in the dataset, the potential for selection bias or statistical manipulations were minimised.

In conclusion, we show that, using baseline clinical and treatment variables, DeepSurv demonstrates superior performance over CPH and RSF for OS prediction in NSCLC patients undergoing RT. We demonstrate that OAR dosimetry can improve survival prediction and use explainable techniques to provide feature importance information for all approaches, particularly facilitating transparency and interpretability of hitherto black-box DL models.

CRediT authorship contribution statement

Joshua R. Astley: Conceptualization, Investigation, Methodology, Software, Validation, Writing – original draft. **James M. Reilly:** Data curation, Investigation, Writing – review & editing. **Stephen Robinson:** Investigation, Resources, Writing – review & editing. **Jim M. Wild:** Funding acquisition, Resources, Supervision, Writing – review & editing. **Matthew Q. Hatton:** Resources, Writing – review & editing. **Bilal A. Tahir:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bilal A. Tahir reports financial support was provided by Yorkshire Cancer Research. Jim M. Wild reports financial support was provided by UKRI Medical Research Council.

Acknowledgements

J.R.A and B.A.T were supported by Yorkshire Cancer Research (S406BT) and Sheffield Hospitals Charity (192032). J.M.W received funding from the Medical Research Council (MR/M008894/1). This study/research is funded by the National Institute for Health and Care Research (NIHR) Sheffield Biomedical Research Centre (NIHR203321). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110084.

References

[1] Bradley JD, Paulus R, Komaki R, Masters G, Blumenschein G, Schild S, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. Lancet Oncol 2015;16:187–99.

[2] Guberina M, Eberhardt W, Stuschke M, Gauler T, Heinzelmann F, Cheufou D, et al. Heart dose exposure as prognostic marker after radiotherapy for resectable stage IIIA/B non-small-cell lung cancer: secondary analysis of a randomized trial. Ann Oncol 2017;28:1084–9.

[3] Shen L, Liu C, Jin J, Han C, Zhou Y, Zheng X, et al. Association of lung and heart dose with survival in patients with non-small cell lung cancer underwent volumetric modulated arc therapy. Cancer Manag Res 2019;11:6091–8.

[4] Cox DR. Regression Models and Life-Tables. J Royal Stat Soc Series B-Methodol 1972;34:187–220.

[5] Hemant I, Udaya BK, Eugene HB, Michael SL. Random survival forests. Ann Appl Stat 2008;2:841–60.

[6] Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. Biostatistics 2006;7:355–73.

[7] Wiegrebe S, Kopper P, Sonabend R, Bender A. Deep Learning for Survival Analysis: A Review. arXiv preprint arXiv:230514961; 2023.

[8] Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Method 2018;18:24.

[9] Bice N, Kirby N, Bahr T, Rasmussen K, Saenz D, Wagner T, et al. Deep learning-based survival analysis for brain metastasis patients with the national cancer database. J Appl Clin Med Phys 2020;21:187–92.

[10] Kim DW, Lee S, Kwon S, Nam W, Cha I-H, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep 2019;9:6994.

[11] Lai Y-H, Chen W-N, Hsu T-C, Lin C, Tsao Y, Wu S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Sci Rep 2020;10:4679.

[12] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135-44.

[13] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Proces Syst 2017;30.

[14] Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP (t): Time-dependent explanations of machine learning survival models. Knowl-Based Syst 2023;262: 110234.

[15] Kovalev MS, Utkin LV, Kasimov EM. SurvLIME: A method for explaining machine learning survival models. Knowl-Based Syst 2020;203:106164.

[16] Robinson SD, Tahir BA, Absalom KAR, Lankathilake A, Das T, Lee C, et al. Radical accelerated radiotherapy for non-small cell lung cancer (NSCLC): A 5-year retrospective review of two dose fractionation schedules. Radiother Oncol 2020; 143:37–43.

[17] Haslett K, Bayman N, Franks K, Groom N, Harden SV, Harris C, et al. Isotoxic Intensity Modulated Radiation Therapy in Stage III Non-Small Cell Lung Cancer: A Feasibility Study. Int J Radiat Oncol Biol Phys 2021;109:1341–8.

[18] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–8.

[19] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 2019;170:W1–33.

[20] Pachón-García C, Hernández-Pérez C, Delicado P, Vilaplana V. SurvLIMEpy: A Python package implementing SurvLIME. arXiv preprint arXiv:230210571; 2023.

[21] Jaworski C, Mariani JA, Wheeler G, Kaye DM. Cardiac complications of thoracic irradiation. J Am Coll Cardiol 2013;61:2319–28.

[22] Arroyo-Hernández M, Maldonado F, Lozano-Ruiz F, Muñoz-Montaño W, Nuñez-Baez M, Arrieta O. Radiation-induced lung injury: current evidence. BMC Pulm Med 2021;21:9.

[23] Spyratos D, Papadaki E, Lampaki S, Kontakiotis T. Chronic obstructive pulmonary disease in patients with lung cancer: prevalence, impact and management challenges. Lung Cancer: Targets and Therapy 2017;8:101–7.

[24] Kimura T, Togami T, Takashima H, Nishiyama Y, Ohkawa M, Nagata Y. Radiation pneumonitis in patients with lung and mediastinal tumours: a retrospective study of risk factors focused on pulmonary emphysema. Br J Radiol 2012;85:135–41.

[25] Fleming C, Cagney DN, O'Keeffe S, Brennan SM, Armstrong JG, McClean B. Normal tissue considerations and dose–volume constraints in the moderately hypofractionated treatment of non-small cell lung cancer. Radiother Oncol 2016; 119:423–31.

[26] Baker S, Fairchild A. Radiation-induced esophagitis in lung cancer. Lung Cancer: Targets and Therapy 2016;7:119–27.

[27] Werner-Wasik M, Paulus R, Curran Jr WJ, Byhardt R. Acute Esophagitis and Late Lung Toxicity in Concurrent Chemoradiotherapy Trials in Patients with Locally Advanced Non–Small-Cell Lung Cancer: Analysis of the Radiation Therapy Oncology Group (RTOG) Database. Clin Lung Cancer 2011;12:245–51.

[28] Saunders M, Dische S, Barrett A, Harvey A, Gibson D, Parmar M. Continuous hyperfractionated accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small-cell lung cancer: a randomised multicentre trial. CHART Steering Committee. Lancet 1997;350:161–5.

[29] McWilliam A, Kennedy J, Hodgson C, Vasquez Osorio E, Faivre-Finn C, van Herk M. Radiation dose to heart base linked with poorer survival in lung cancer patients. Eur J Cancer 2017;85:106–13.

[30] Sun W, Jiang M, Dang J, Chang P, Yin FF. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiat Oncol 2018;13:197.

[31] Jin L, Zhao Q, Fu S, Cao F, Hou B, Ma J. Development and validation of machine learning models to predict survival of patients with resected stage-III NSCLC. Front Oncol 2023;13:1092478.

[32] Lee SH, Geng H, Arnold J, Caruana R, Fan Y, Rosen MA, et al. Interpretable Machine Learning for Choosing Radiation Dose-volume Constraints on Cardio-pulmonary Substructures Associated with Overall Survival in NRG Oncology RTOG 0617. Int J Radiat Oncol Biol Phys 2023.

[33] Meyers DE, Pasternak M, Dolter S, Grosjean HAI, Lim C, Stukalin I, et al. Impact of performance status on survival outcomes and health care utilization in patients with advanced non–small cell lung cancer treated with immune checkpoint inhibitors. J Clin Oncol 2022;40:9053.

[34] Käsmann L, Taugner J, Eze C, Roengvoraphoj O, Dantes M, Gennen K, et al. Performance Status and Its Changes Predict Outcome for Patients With Inoperable Stage III NSCLC Undergoing Multimodal Treatment. Anticancer Res 2019;39: 5077–81.

[35] Khan M, Ambady P, Kimbrough D, Shoemaker T, Terezakis S, Blakeley J, et al. Radiation-Induced Myelitis: Initial and Follow-Up MRI and Clinical Features in Patients at a Single Tertiary Care Institution during 20 Years. AJNR Am J Neuroradiol 2018;39:1576–81.