Clinical and Translational Radiation Oncology 22 (2020) 69–75

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.elsevier.com/locate/ctro

Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation

José Marcio Luna [a,1,2,*], Hann-Hsiang Chao [b,1], Russel T. Shinohara [c,2], Lyle H. Ungar [d], Keith A. Cengel [a], Daniel A. Pryma [e], Chidambaram Chinniah [f], Abigail T. Berman [a], Sharyn I. Katz [e], Despina Kontos [e], Charles B. Simone II [g], Eric S. Diffenderfer [a]

[a] Department of Radiation Oncology, University of Pennsylvania, Perelman Center for Advanced Medicine, 3400 Civic Center Blvd, Philadelphia, PA 19104, United States
[b] Department of Radiation Oncology, Hunter Holmes McGuire Veterans Affairs Medical Center, 1201 Broad Rock Blvd, Richmond, VA 23249, United States
[c] Department of Biostatistics and Epidemiology, University of Pennsylvania, 423 Guardian Dr, Philadelphia, PA 19104, United States
[d] Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St, Philadelphia, PA 19104, United States
[e] Department of Radiology, University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104, United States
[f] Albany Medical College, 43 New Scotland Ave, Albany, NY 12208, United States
[g] Department of Radiation Oncology, New York Proton Center, 225 East 126th St, New York, NY 10035, United States

ARTICLE INFO

Article history:
Received 16 December 2019
Revised 17 March 2020
Accepted 21 March 2020
Available online 24 March 2020

Keywords:
Radiation esophagitis
Machine learning
Non-small cell lung cancer
Chemoradiation
Radiation-induced toxicity
Intensity-modulated radiation therapy
Proton beam therapy

ABSTRACT

Background and Purpose: Radiation esophagitis is a clinically important toxicity seen with treatment for locally-advanced non-small cell lung cancer. There is considerable disagreement among prior studies in identifying predictors of radiation esophagitis. We apply machine learning algorithms to identify factors contributing to the development of radiation esophagitis to uncover previously unidentified criteria and more robust dosimetric factors.

Materials and Methods: We used machine learning approaches to identify predictors of grade $\geq$ 3 radiation esophagitis in a cohort of 202 consecutive locally-advanced non-small cell lung cancer patients treated with definitive chemoradiation from 2008 to 2016. We evaluated 35 clinical features per patient grouped into risk factors, comorbidities, imaging, stage, histology, radiotherapy, chemotherapy and dosimetry. Univariate and multivariate analyses were performed using a panel of 11 machine learning algorithms combined with predictive power assessments.

Results: All patients were treated to a median dose of 66.6 Gy at 1.8 Gy per fraction using photon (89.6%) and proton (10.4%) beam therapy, most often with concurrent chemotherapy (86.6%). 11.4% of patients developed grade $\geq$ 3 radiation esophagitis. On univariate analysis, no individual feature was found to predict radiation esophagitis (AUC range 0.45–0.55, $p \geq 0.07$). In multivariate analysis, all machine learning algorithms exhibited poor predictive performance (AUC range 0.46–0.56, $p \geq 0.07$).

Conclusions: Contemporary machine learning algorithms applied to our modern, relatively large institutional cohort could not identify any reliable predictors of grade $\geq$ 3 radiation esophagitis. Additional patients are needed, and novel patient-specific and treatment characteristics should be investigated to develop clinically meaningful methods to mitigate this survival altering toxicity.

1. Introduction

Severe radiation esophagitis is a clinically important toxicity that frequently arises during the treatment of locally advanced non-small cell lung cancer (LA-NSCLC) [1–3]. Radiation esophagitis acutely can be present as dysphagia, odynophagia, sternal or epigastric chest pain, or spasms, which can directly influence patient

[*] Corresponding author at: Bldg 421, SCTR 8-130, 3400 Civic Center Blvd, Philadelphia, PA 19104, United States.

E-mail address: jose.luna@pennmedicine.upenn.edu (J.M. Luna).

[1] These authors contributed equally to this work.

[2] These authors were responsible for the statistical analyses.

https://doi.org/10.1016/j.ctro.2020.03.007

quality of life [4], or as acute or late esophageal bleeding, perforation, or fistulas, which can be life threatening [5].

The estimated incidence of this toxicity ranges from 7 to 25% in patients receiving standard of care definitive chemoradiation [6–9] and development of high-grade ($\geq$3) radiation esophagitis can necessitate interventions such as analgesic medications, treatment delays/breaks, hospitalizations, and permanent feeding tube dependence [10]. Prior efforts aimed at improving the survival of LA-NSCLC using radiation dose-escalation were unsuccessful likely in part due to the dose limiting toxicity of radiation esophagitis. In prior multi-institutional randomized clinical trials, high-grade radiation esophagitis is shown to negatively affect overall survival (OS) [7,11], highlighting the importance of mitigating this toxicity.

Prior attempts at identifying predictors of esophagitis have identified the importance of factors such as concurrent chemotherapy, radiation dose intensification, and dosimetric/volumetric factors related to the esophagus itself, but there are conflicting data on the predictors, especially in terms of dose constraints [1,9,12–18]. As such, currently there is no consensus for predicting and preventing radiation esophagitis regarding optimal thresholds for volume criteria, dose-volume criteria, radiation treatment modality, or the comparative importance of these factors. This study aims to employ machine learning techniques to identify the critical predictors of esophageal toxicity and their comparative importance in order to inform clinical decision making. Here, we analyze 35 continuous and categorical variables drawn from previous literature as predictors of grade $\geq$ 3 radiation esophagitis on a large institutional cohort of 202 consecutively treated LA-NSCLC patients. We apply three variants of a panel of 11 machine learning techniques to robustly identify the important predictive factors in the development of grade $\geq$ 3 radiation esophagitis.

2. Methods and materials

2.1. Patient cohort

With institutional review board approval (Penn IRB protocol #832329), we identified a cohort of 202 consecutive patients with histologically confirmed Stage II-III LA-NSCLC (AJCC 7th Edition) treated at our institution with sequential or concurrent chemoradiation with platinum-containing regimens between 2008 and 2016. Patients received treatment using either proton beam therapy (PBT) or intensity-modulated radiation treatment (IMRT) with x-rays. Radiation esophagitis was graded according to the common terminology criteria for adverse events (CTCAEv4.0).

2.2. Feature definition

In this study, we analyzed a set of 35 predefined continuous and categorical features, including variables previously reported in the literature as strong predictors of grade $\geq$ 3 radiation esophagitis. The categorical features were ethnicity, pre-treatment ECOG, 3-month post-radiotherapy (RT) ECOG performance status, AJCC clinical stage grouping, T Stage, N Stage, radiation treatment modality (photon or proton), concurrent vs. sequential chemotherapy, specific chemotherapy agents used, tumor grade, and sex. The continuous features were smoking pack-years (pack-year), body mass index (BMI) age at diagnosis, primary tumor size, pulmonary function test (PFT) pre-bronchodilator, DLCO (% predicted), PFT pre-bronchodilator FEV1 (L), radiation total dose (total dose), radiation fraction size, number of radiation fractions (nr. fractions), mean esophagus dose (eso mean), maximum esophagus dose (eso max), eso V40, eso V50, eso V60, mean lung dose (lung mean), lung V5, lung V10, lung V20, mean heart dose (heart mean), heart V5, heart V30, heart V50 and heart V60. For PBT, the dosimetric indices were calculated using the proton convolution superposition algorithm (Varian Medical Systems, Palo Alto, CA, USA), and for IMRT dose calculations with heterogeneity corrections were performed using the analytical anisotropic algorithm (photons). This set of dose parameters and clinical features was thoroughly discussed and selected by three highly experienced board-certified thoracic radiation oncologists (CBS, ATB, KAC) at our institution based on their expertise, best clinical practice, and current national treatment guidelines.

2.3. Missing values imputation

Missing values were imputed using trimmed scores regression (TSR), a method that fits principal component analysis (PCA) models iteratively thus exploiting the statistical relationship among features [19]. This imputation is based on the first four principal components, which for this cohort explain 95.84% of the variance of the data. More details about the selection of TSR imputation in Section S3 of the SI Appendix.

2.4. Univariate analysis

Based on the two labeled classes (esophagitis/non-esophagitis) in our cohort, we performed a Wilcoxon rank-sum test for each continuous predictor as well as a χ^2 test for each categorical predictor. The statistical significance (p-value) of the separation between the two classes by each predictor was estimated. Due to Bonferroni correction of a 5% family-wise error rate, the significance level $\alpha = 0.002$ was used for multiple comparisons. The average performance indexes, specifically, balanced accuracy (BACC) [20], the receiver operating characteristic (ROC), the area under the ROC curve (AUC) were estimated. BACC is defined as the average between sensitivity and specificity, and commonly used to calculate performance in two-class imbalanced domains. 95% confidence intervals of the average performance measurements (i.e., BACC and AUC) were calculated using cross-validated estimates as bootstrapped samples and using the standard t-distribution-based approximation. A total of 500,000 bootstrap replicates were used to estimate the confidence intervals. All the analyses were implemented using the Statistics and Machine Learning Toolbox of Matlab R2018b® (MathWorks, Santa Clara, CA, USA) [21]. Additionally, Pearson correlation coefficients were calculated to assess possible confounders associated with radiation esophagitis.

2.5. Multivariate analysis

To assess the combined capacity of prediction of the features, we used a set of diverse statistical tools including long-existing methods such as logistic regression [22], elastic net, k-nearest neighbors (k-NN) and linear and quadratic discriminants [23] to more sophisticated methods such as linear, quadratic and Gaussian support vector machines (SVM) [24], classification and regression trees (CART) [25], Random Forest [26] and boosted trees (RUSBoost) [27]. All experiments were performed using nested resampling as shown in Fig. S1. We implemented stratified 5-fold cross-validation for the internal resampling, where the validation set, was used for hyperparameter tuning and feature selection using grid search to maximize BACC. We also used stratified 5-fold cross-validation for the external resampling. The test set in the external resampling (Fig. S1), also known as hold-out set, was used for performance estimation of the model, i.e., BACC and AUC. Same as univariate analysis, 95% confidence intervals of the average performance measurements were calculated using cross-validated estimates as bootstrapped samples using the standard t-distribution-based approximation with 500,000 bootstrap replicates. We assured that the observations used in hyperparameter optimization never appear in the external resampling test set, thus reducing model overfitting. The list of hyperparameters tuned per each implemented algorithm is shown in Table S1. This analysis corresponds to the development and validation of a predictive model using resampling or analysis type 1b as specified in Collins et al, 2015 [28]. The implemented stages of nested resampling are illustrated in the workflow shown in Fig. 1. In the internal resampling for each machine learning algorithm, one model is built per fold, for a total of five models. Then, the hyperparameters of the model with the highest BACC calculated through grid search, are selected and the performance of such model (i.e., BACC and AUC) is subsequently assessed on the respective test set during external resampling.

For further exploration of the prediction power of the algorithms, three variants of the experiments were proposed:

1. Evaluation of predictive power using all 35 predictors.
2. Predictive power assessment using backward sequential feature selection (BSFS).
3. Predictive power assessment using synthetic minority oversampling technique (SMOTE) [29] and BSFS.

SMOTE is a method that combines the under-sampling of the majority class and the over-sampling of the minority class by creating synthetic minority class examples. This increases the sensitivity of the classifiers to the minority class [29]. In the experiments where oversampling was implemented, SMOTE was performed in the internal resampling only, specifically in the training set of each internal fold. Moreover, BSFS was performed in the internal resampling, in the experiments in which feature selection was implemented. The p-values associated with the predictive performance of the different algorithms were calculated using the Wilcoxon rank-sum test with a significance level $\alpha = 0.002$, due to Bonferroni correction of a 5% error rate, considering the three variants of the 11 algorithms. For each of the machine learning algorithms, we calculate their predictions using the test sets in the outer resampling. Following the Wilcoxon rank sum test procedure, we compare the distribution of the estimated predictions of each algorithm between the two labeled classes (esophagitis/non-esophagitis).

3. Results

3.1. Patient characteristics and outcomes

Characteristics of the 202 consecutive patients with adenocarcinoma who were treated at our center with chemoradiation and included in the current analysis are provided in Tables 1 and 2. Patients were treated homogeneously to a median dose of 66.6 Gy at 1.8 Gy per fraction (range 64.8–66.6 Gy at 1.8 Gy per fraction). The median age of the cohort was 64 years (range 56–73). Radiation was mainly delivered with IMRT (89.6%), with a minority receiving proton beam therapy (10.4%). Overall, 86.6% of patients received concurrent chemotherapy, with a carboplatin-based doublet combination (51.5%) being the most common regimen, followed by cisplatin-based doublet (34.7%).

At a median follow-up of 22.6 months (1–88 month range), patients had a median OS of 23.5 months, 1-year OS of 75.0%, 2-year OS of 49.0%, and 5-year OS of 12.0%, all calculated from a Kaplan-Meier plot. Within the cohort, 23 patients (11.4%) developed grade ≥ 3 radiation esophagitis.

3.2. Univariate analysis

Dosimetric parameters of the same organ at risk (e.g. heart, lung, and esophagus) were found to be strongly positively correlated to each other and also showed weaker positive correlations with neighboring anatomic organs (e.g. lung-heart, lung-esophagus) (Fig. 2). The univariate analysis showed that no individual features can predict grade ≥ 3 radiation esophagitis, with median AUC = 0.49 (range 0.45–0.55) and $p \geq 0.07$ across all 35 features (Table 3).

3.3. Multivariate analysis

The predictive power using AUCs, as well as the associated p-values and the optimal BACC for all classifiers are summarized in Table 4. In the first experiment where we trained all 11 algorithms

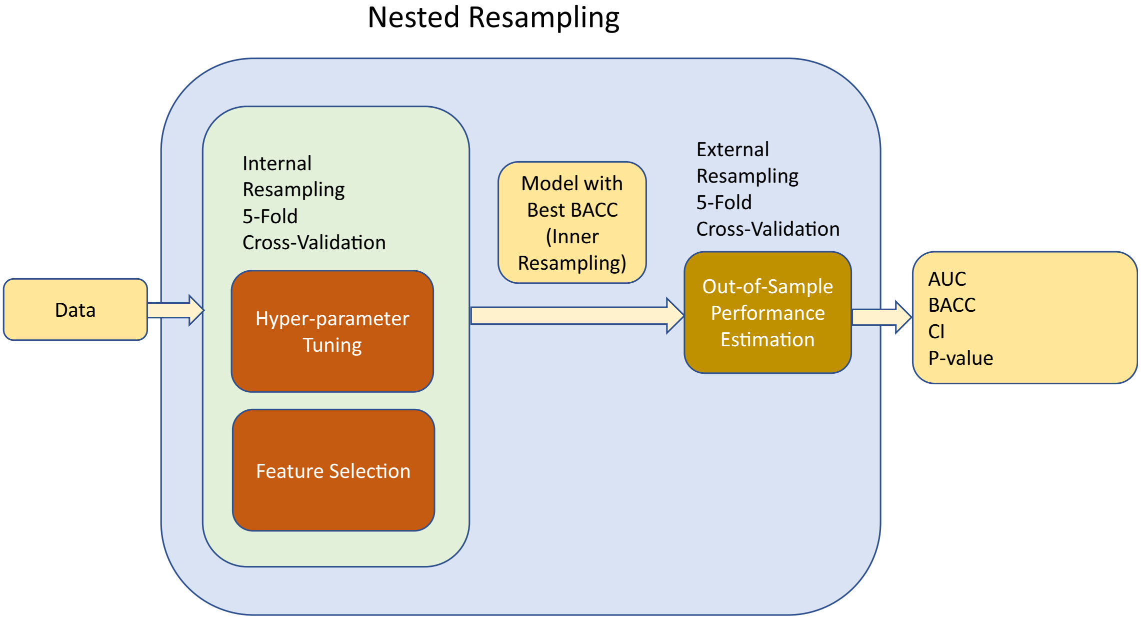

Fig. 1. Multivariate analysis workflow. Diagram illustrating the workflow by which the input data undergoes stepwise resampling to estimate model performance for prediction of radiation esophagitis.

using the complete set of 35 predictors, none of the algorithms combining the effect of all available features could predict grade $\geq$ 3 radiation esophagitis with median AUC = 0.50 (range 0.45–0.54) and p $\geq$ 0.09 across all the algorithms (upper third of Table 4). In the second experiment where BSFS in the internal resampling was implemented, a median AUC = 0.52 (range 0.49–0.56) and p $\geq$ 0.25 across all machine learning algorithms show that the algorithms are unable to perform better than a random classifier (middle third of Table 4). It evidences the lack of capacity of the current combined features to separate grade $\geq$ 3 radiation esophagitis even when counteracting confounding through feature selection. BSFS was chosen over forward sequential feature selection (FSFS) due to its superior predictive performance using logistic regression in our cohort (Fig. S2). In the last set of experiments, where SMOTE was implemented, a median AUC = 0.49 (range 0.45–0.52) and p $\geq$ 0.07, show that no algorithm significantly predicted the two classes.

4. Discussion

Prior attempts to identify predictors of radiation esophagitis have utilized various methodologies, with many of these reports conducted in a pre-IMRT era, resulting in conflicting data regarding the relative importance of these factors [9,12–14,16–18,30,31] as illustrated in Table S1. We sought to apply machine learning algorithms to a contemporary, curated patient cohort treated relatively homogenously with modern radiotherapy techniques in order to examine, validate, and rank factors suggested to predict for high-grade esophagitis. Here, we specifically analyzed predictors of grade $\geq$ 3 radiation esophagitis, a particularly important toxicity and grade given its association with considerably worse OS as reported on the clinical trial of the Radiation Therapy Oncology Group (RTOG) 0617 [7]. Interestingly, we found that when using a combination of machine learning methodologies coupled with resampling techniques to reduce confounding from overfitting, no single feature reliably predicts grade $\geq$ 3 esophagitis in our analysis.

Table 1
Summary of categorical patient characteristics. Description of clinical characteristics of the cohort with their respective categorization and percentages.

Categorical Predictors	Classes	Number of Patients	(%)
Sex	Male	90	44.6
	Female	112	55.4
Smoking History	Former	136	67.3
	Current	26	12.9
	Never	17	8.4
	Not Available	23	11.4
Ethnicity	White	137	67.8
	Black	47	23.3
	Asian	4	2.0
	Other	14	6.9
Pre Treatment ECOG Perform. Status	0	77	38.1
	1	55	27.2
	2	14	6.9
	3	2	1.0
	4	2	1.0
	Not Recorded	52	25.7
Stage Grouping	IIB	1	0.5
	IIIA	120	58.9
	IIIB	81	40.6
Tumor Stage	Tx	15	7.4
	T1	51	25.2
	T2	63	31.2
	T3	32	15.8
	T4	41	20.3
Nodal Stage	Nx	7	3.5
	N0	9	4.5
	N1	12	5.9
	N2	126	62.4
	N3	48	23.8
Histology	Adenocarcinoma	202	100.0
Radiation Modality	Photon (IMRT)	181	89.6
	Proton	21	10.4
Chemotherapy	Concurrent	176	86.6
	Sequential	21	10.4
	None	5	3.0
Chemotherapy Agents	Carboplatin-based Doublet	104	51.5
	Cisplatin-based Doublet	70	34.7
	Platinum-based Triplet	6	3.0
	Single Agent	2	1.0
	Other	20	9.9

Table 2
Summary of numerical patient characteristics. Description of numerical characteristics of the cohort with their respective median and interquartile ranges.

Continuous Predictors	Median	Range ‡
Age (yr)	64	(56–73)
Pack-Year (current/former smokers)	35	(14.5–50)
BMI (kg/m^2)	26.0	(23.0–30.0)
Radiation Dose Delivered (Gy)	66.6	(64.8–66.6)
Dose per fraction (Gy)	1.8	(1.8–1.8)
Esophagus Mean Dose (Gy)	24.5	(18.3–31.9)
Esophagus Maximum Dose (Gy)	69.4	(65.4–72.4)

‡ Interquartile range.

Although we found that eso mean and dosimetric factors that correlated strongly with esophageal dose (Heart Mean, Heart V30) ranked highly in feature importance (Table 3), none crossed the AUC threshold in our study to be deemed a reliable predictor. This is in contrast to previously published retrospective [12,32], prospective [16,33] and randomized [34] studies showing associations between esophageal toxicity and predictors including age, tumor nodal stage, concurrent chemotherapy and BMI. It also contrasts with dosimetric factors including eso mean, eso max, as well as eso V20, eso V35, eso V60 which have been analyzed retrospectively in [35], and using a mixture of retrospective and prospective collected datasets in [9]. It is important to note that we specifically examined grade $\geq$ 3 RE, whereas much of the prior literature has examined grade $\geq$ 2 esophagitis, a less clinically impactful toxicity [36,37].

In our dataset, we observed grade $\geq$ 3 radiation esophagitis in 23 of 202 patients (11.4%), which is lower than historically observed rates of >15%[3,10,35,37]. The comparatively low radiation esophagitis rates we report here may reflect treatment improvements over eras reflecting improved symptom prevention and proactive care, as well as advanced radiation treatment modalities. Importantly, to our knowledge, our series is the first machine learning analysis to include a cohort of patients treated with proton beam therapy, which may also result in lower than expected radiation esophagitis rates, as has been reported in locally advanced lung cancer prospective population treated with proton therapy [38]. Additionally, the lower than expected toxicity rates may contribute to a lack of reliable predictors in our models due to insufficient radiation esophagitis events. By comparison, other studies reporting robust predictors for grade $\geq$ 2 radiation esophagitis observed toxicity rates upwards of 50% [9,10,35].

Another inherent challenge in using machine learning tools to identify predictive factors is falsely identifying significant factors due to overfitting. We also sought to enhance the robustness of our machine learning models through the use of sequential feature selection and resampling using BSFS and SMOTE. We further attribute the difference in results between our current models and the prior literature to a combination of the different toxicity endpoint assessed (grade $\geq$ 3 vs. grade $\geq$ 2), variance in radiation techniques, and implementation of resampling. Previously published

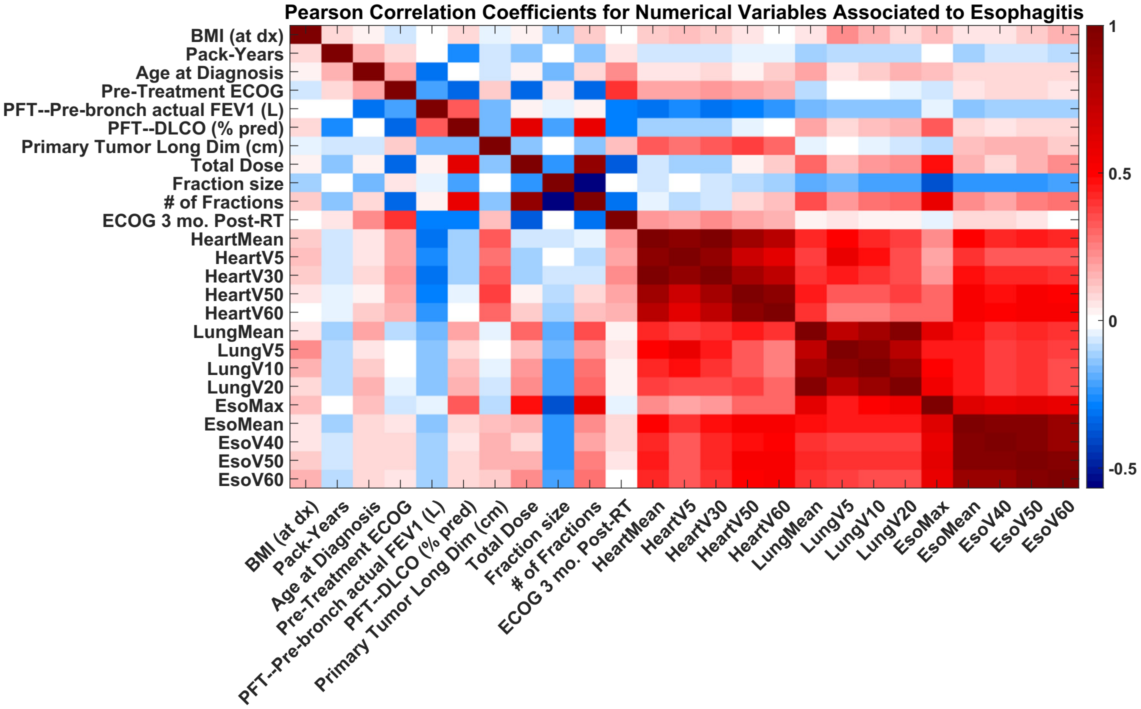

Fig. 2. Feature correlation heat map. Heat map, illustrating the Pearson correlation between the continuous features under study.

Table 3
Univariate analysis. Predictive performance for individual features using AUC analysis with their respective significance using Wilcoxon rank-sum test (continuous features) and χ^2 (categorical features). None of the features can predict grade $\geq$ 3 RE using Bonferroni correction ($\alpha = 0.002$) for multiple comparison.

Feature	AUC §	P-value
T Stage	0.41 (0.28,0.55)	0.07
Lung V20	0.39 (0.27,0.52)	0.09
BMI	0.61 (0.48,0.72)	0.09
Pack Years	0.40 (0.29,0.52)	0.12
Concurrent v Sequential	0.57 (0.55,0.60)	0.15
Eso V60	0.59 (0.47,0.70)	0.16
Lung Mean	0.41 (0.30,0.54)	0.16
Total Dose	0.42 (0.31,0.55)	0.18
Heart Mean	0.58 (0.43,0.71)	0.23
Agents Drugs	0.43 (0.33,0.54)	0.25
Heart V30	0.57 (0.43,0.70)	0.28
Pre Treatment ECOG	0.56 (0.43,0.68)	0.31
Sex	0.56 (0.44,0.65)	0.32
Eso Max	0.44 (0.33,0.56)	0.33
Eso V50	0.56 (0.44,0.67)	0.34
Ethnicity	0.45 (0.37,0.58)	0.34
Eso V40	0.56 (0.43,0.67)	0.38
N Stage	0.55 (0.45,0.62)	0.38
Heart V5	0.55 (0.41,0.69)	0.42
Heart V60	0.55 (0.41,0.67)	0.43
Best CS AJCC Stage	0.43 (0.32,0.54)	0.44
Heart V50	0.55 (0.41,0.67)	0.48
Age at Diagnosis	0.46 (0.36,0.57)	0.52
Lung V10	0.46 (0.33,0.58)	0.52
Nr of Fractions	0.46 (0.37,0.59)	0.55
Eso Mean	0.54 (0.41,0.66)	0.55
Grade Differentiation	0.45 (0.36,0.56)	0.56
Fraction Size	0.53 (0.41,0.61)	0.57
PFT DLCO pred	0.47 (0.36,0.59)	0.63
Linac	0.51 (0.46,0.62)	0.66
Proton	0.49 (0.38,0.54)	0.66
Lung V5	0.47 (0.34,0.60)	0.67
PFT Pre Bronch Actual FEV1 L	0.52 (0.40,0.63)	0.76
Primary Tumor Long Dim cm	0.48 (0.35,0.61)	0.78
ECOG 3 mo Post-RT	0.49 (0.37,0.61)	0.85

§ Estimate with 95% confidence interval.

reports using models developed using a median cohort size of 141 (well below our actual cohort), and median rate of radiation esophagitis of 11.1%, used logistic regression approaches [1,9,12,17,31,32,34,36,37,39], while some other studies would fit Lyman-Kutcher-Burman (LKB) models, both using pre-IMRT era cohorts [14,33], which may not reflect the current risks of radiation-induced toxicity in a contemporary setting. A more recent study used lasso regularization using a smaller cohort of 94 patients, where 16% developed radiation esophagitis [35]. Finally, some studies with considerably larger cohorts are limited to conformal radiotherapy patients and/or a rather small number of IMRT patients [9,14]. To the best of our knowledge, this is the first study with a relatively large cohort using sophisticated machine learning techniques to identify predictors of grade $\geq$ 3 radiation esophagitis.

Radiation esophagitis has been a challenging entity to predict, which is reflected in the variance in esophageal dosimetric constraints employed in recent national prospective randomized trials on LA-NSCLC. Among the two most recent NRG oncology phase III randomized trials for LA-NSCLC, RTOG 0617 [7] recommended constraint of an eso mean below 34 Gy and to record the eso V60 without required dose constraint, whereas RTOG 1308 [40], the currently enrolling prospective trial comparing proton vs. photon radiation therapy, set a per protocol constraint of an eso max of 74 Gy to $\leq$1 cc of the partial circumference while retaining none of the earlier constraints from RTOG 0617. This is in contrast to the pulmonary constraints, which have remained relatively constant over these trials.

An earlier meta-analysis focusing on radiation esophagitis development [9] illustrates some of the potential challenges in identifying reliable predictors. Similar to our analysis, many clinical and dosimetric factors are found to be associated with radiation esophagitis toxicity, with more features associated with grade $\geq$ 2 than grade $\geq$ 3 radiation esophagitis. However, when these features are used as predictors, they by and large perform poorly with C-statistics below 0.6. Interestingly, eso V60 was identified as a

Table 4
Multivariate analysis. Combined predictive performance of features using 11 statistical models with three variants namely, a) 35 handcrafted features, b) BSFS and c) BSFS and SMOTE. None of the models can predict grade $\geq$ 3 RE using Bonferroni correction ($\alpha = 0.002$) for multiple comparison.

Experiment	Algorithm	AUC §	BACC §	P-value
All 35 Features	Logistic Regression	0.58 (0.27,0.88)	0.58 (0.29,0.87)	0.09
	Linear Discriminant	0.57 (0.21,0.93)	0.57 (0.25,0.89)	0.30
	Linear SVM	0.56 (0.22,0.90)	0.49 (0.38,0.60)	0.50
	Elastic Net	0.52 (0.17,0.87)	0.47 (0.28,0.66)	0.88
	RUSBoost	0.52 (0.07,0.96)	0.56 (0.23,0.88)	0.62
	k-NN	0.50 (0.20,0.79)	0.53 (0.27,0.79)	0.72
	Quadratic SVM	0.49 (0.08,0.89)	0.53 (0.19,0.88)	0.74
	Random Forest	0.46 (0.10,0.82)	0.50 (0.50,0.50)	0.55
	Quadratic Discriminant	0.45 (0.09,0.80)	0.48 (0.14,0.82)	0.27
	CART	0.44 (0.17,0.71)	0.47 (0.33,0.60)	0.32
	Gaussian SVM	0.40 (0.03,0.77)	0.50 (0.48,0.51)	0.12
BSFS	Logistic Regression	0.61 (0.41,0.81)	0.54 (0.31,0.78)	0.26
	Linear Discriminant	0.59 (0.30,0.88)	0.52 (0.34,0.70)	0.25
	Linear SVM	0.57 (0.50,0.64)	0.50 (0.46,0.54)	0.54
	Random Forest	0.56 (0.22,0.90)	0.53 (0.35,0.71)	0.35
	k-NN	0.53 (0.19,0.86)	0.56 (0.25,0.87)	0.71
	Elastic Net	0.52 (0.17,0.87)	0.47 (0.28,0.66)	0.88
	Quadratic SVM	0.50 (0.14,0.85)	0.50 (0.23,0.78)	0.84
	RUSBoost	0.49 (0.05,0.93)	0.49 (0.18,0.81)	0.76
	Quadratic Discriminant	0.48 (0.09,0.88)	0.50 (0.23,0.77)	0.66
	Gaussian SVM	0.46 (0.13,0.78)	0.52 (0.43,0.61)	0.73
	CART	0.40 (0.20,0.60)	0.46 (0.41,0.52)	0.25
BSFS and SMOTE	Elastic Net	0.61 (0.16,1.00)	0.63 (0.24,1.00)	0.07
	Linear Discriminant	0.58 (0.17,0.98)	0.55 (0.21,0.89)	0.39
	Logistic Regression	0.55 (0.15,0.95)	0.54 (0.18,0.90)	0.42
	Linear SVM	0.50 (0.14,0.86)	0.51 (0.34,0.68)	0.85
	Quadratic SVM	0.49 (0.12,0.86)	0.52 (0.13,0.90)	0.92
	RUSBoost	0.49 (0.12,0.85)	0.50 (0.18,0.82)	0.73
	Random Forest	0.48 (0.12,0.83)	0.53 (0.26,0.80)	0.73
	k-NN	0.47 (0.10,0.85)	0.51 (0.19,0.83)	0.61
	Gaussian SVM	0.43 (0.02,0.84)	0.49 (0.22,0.76)	0.33
	CART	0.42 (0.01,0.83)	0.48 (0.21,0.76)	0.35
	Quadratic Discriminant	0.41 (0.12,0.69)	0.48 (0.45,0.51)	0.54

§ Estimate with 95% confidence interval.

reliable predictor of grade $\geq$ 3 radiation esophagitis in [9]. This series does represent an older cohort of patients treated from 1993 to 2011, which may contribute to some of the discrepant findings. Furthermore, none of the eso V40, V50 nor V60 were identified as an important predictor in our study (see Table 3).

Given the result from our machine learning analysis that none of the 35 features analyzed performed better than a random classifier, this suggests that our currently utilized clinical, demographic, and dosimetric features could be inadequate to reliably predict radiation esophagitis. One can conclude that we are not currently collecting and capturing the appropriate features to allow a machine learning workflow to predict grade $\geq$ 3 radiation esophagitis, as we were successfully able to do when using machine learning to predict for pneumonitis in LA-NSCLC [41] and chest wall toxicity in early stage NSCLC [42]. As such, we encourage other investigators to explore and develop new markers directed at this toxicity. This may include more widespread utilization of biomarkers or composite features to generate the appropriate power and granularity to adequately capture radiation esophagitis.

In the recent work of Bahn and Alber [43], the authors assume a unimodal beta distribution of the output of the normal tissue complication probability (NTCP), and using Monte Carlo simulations state that a cohort size of N = 300 is necessary to be powered to detect a small difference of 0.1 between two AUCs. Our current cohort size N = 202 does not fulfill this suggested sample size for AUC comparison in weak model settings. It does, however, meet the sample size recommendations for detecting models with medium predictive performance (as defined in [43], AUC = 0.69). In summary, our findings encourage the incorporation of novel predictors of acute esophagitis in our future research agenda, as well as the prospective increase of the cohort size as more patient information becomes available at our institution. This single institution analysis, however, does allow us comparative uniformity in the patient population and minimizes potential heterogeneity in the assessed population. It is also worth noting that our cohort is the largest used in a modern, IMRT analysis of grade > 3 RE in stage II-III NSCLC patients performed to date.

5. Conclusions

From our analysis, we conclude that current predictors for high-grade radiation esophagitis are unreliable and that continued investigation is necessary to develop clinically useful metrics for prevention of this detrimental toxicity that is associated with overall survival. Reporting and identifying more robust variables will be critically important for future study. Clinicians should employ individualized patient-centered decision making in terms of treatment regimens and toxicity mitigation until reliable radiation esophagitis predictors can be identified.

Acknowledgements

This work was partially supported by an award granted by the Emerson Collective Research Fund. Special thanks to Grace Ma for her assistance extracting clinical information for the cohort.

Data statement

Should a researcher wish to investigate this dataset, efforts would be made to establish a material transfer agreement and IRB approval to share the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2020.03.007.

References

[1] Werner-Wasik M, Pequignot E, Leeper D, Hauck W, Curran W. Predictors of severe esophagitis include use of concurrent chemotherapy, but not the length of irradiated esophagus: a multivariate analysis of patients with lung cancer treated with nonoperative therapy. Int J Radiat Oncol Biol Phys 2000;48:689–96.

[2] Rose J, Rodrigues G, Yaremko B, Lock M, Souza DD. Systematic review of dose – volume parameters in the prediction of esophagitis in thoracic radiotherapy. Radiother Oncol 2009;91:282–7. https://doi.org/10.1016/j.radonc.2008.09.010.

[3] Werner-wasik M, Paulus R, Curran 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. https://doi.org/10.1016/j.cllc.2011.03.026.

[4] Movsas B, Hu C, Sloan J, Bradley J, Komaki R, Masters G, et al. Quality of life analysis of a radiation dose-escalation study of patients with non-small-cell lung cancer a secondary analysis of the radiation Therapy Oncology Group 0617 randomized clinical trial. JAMA Oncol 2016;2:359–67. https://doi.org/10.1001/jamaoncol.2015.3969.

[5] Simone 2nd CB. Thoracic radiation normal tissue injury. Semin Radiat Oncol 2017;27:370–7. https://doi.org/10.1016/j.semradonc.2017.04.009.

[6] Auperin A, Le Pechoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol 2010;28:2181–90.

[7] 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 randomized two-by-two factorial ph. Lancet Oncol 2015;16:187–99. https://doi.org/10.1016/S1470-2045(14)71207-0.

[8] Curran WJ, Paulus R, Langer CJ, Komaki R, Lee JS, Hauser S, et al. Sequential vs concurrent chemoradiation for stage III non–small cell lung cancer: randomized phase III trial RTOG 9410. J Natl Cancer Inst 2011;103:1452–60. https://doi.org/10.1093/jnci/djr325.

[9] Palma DA, Senan S, Oberije C, Belderbos J, De Dios NR, Bradley JD, et al. Predicting esophagitis after chemoradiation therapy for non-small cell lung cancer: an individual patient data meta-analysis. Int J Radiat Oncol Biol Phys 2013;87:690–6. https://doi.org/10.1016/j.ijrobp.2013.07.029.

[10] Werner-wasik M. Treatment-related esophagitis. Semin Oncol 2005:60–6. https://doi.org/10.1053/j.seminoncol.2005.03.011.

[11] Machtay M, Hsu C, Komaki R, Sause WT, Swann S, Langer CJ, et al. Effect of overall treatment time on outcomes after concurrent chemoradiation for locally advanced non-small cell lung carcinoma: analysis of the radiation therapy oncology group (RTOG) experience. Int J Radiat Oncol Biol Phys 2005;63:667–71. https://doi.org/10.1016/j.ijrobp.2005.03.037.

[12] Bradley J, Deasy JO, Betzen S, El Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. Int J Radiat Oncol Biol Phys 2004;58:1106–13. https://doi.org/10.1016/j.ijrobp.2003.09.080.

[13] Bradley J, Movsas B. Radiation esophagitis: predictive factors and preventive strategies. Semin Radiat Oncol 2004;14:280–6. https://doi.org/10.1053/j.semradonc.2004.06.003.

[14] Gomez DR, Tucker SL, Martel MK, Mohan R, Balter PA, Lopez L, et al. Predictors of high-grade esophagitis after definitive three-dimensional conformal therapy, intensity-modulated radiation therapy, or proton beam therapy for non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2012;84:1010–6. https://doi.org/10.1016/j.ijrobp.2012.01.071.

[15] Mehmood Q, Sun A, Becker N, Higgins J, Marshall A, Le LW, et al. Predicting radiation esophagitis using 18 F-FDG PET during chemoradiotherapy for locally advanced non–small cell lung cancer. J Thorac Oncol 2015;11:213–21. https://doi.org/10.1016/j.jtho.2015.10.006.

[16] Patel AB, Edelman MJ, Kwok Y, Krasna MJ, Suntharalingam M. Predictors of acute esophagitis in patients with non-small cell lung carcinoma treated with concurrent chemotherapy and hyperfractionated radiotherapy followed by surgery. Int J Radiat Oncol Biol Phys 2004;60:1106–12. https://doi.org/10.1016/j.ijrobp.2004.04.051.

[17] Maguire PD, Sibley GS, Zhou S-M, Jamieson TA, Light KL, Antoine PA, et al. Clinical and dosimetric predictors of radiation-induced esophageal toxicity. Int J Radiat Oncol Biol Phys 1999;45:97–103.

[18] Singh AK, Lockett MA, Bradley JD. Predictors of radiation-induced esophageal toxicity in patients with non-small cell lung cancer treated with three-dimensional conformal radiotherapy. Int J Radiat Oncol Biol Phys 2003;55:337–41.

[19] Folch-Fortuny A, Arteaga F, Ferrer A. PCA model building with missing data: new proposals and a comparative study. Chemom Intell Lab Syst 2015;146:77–88. https://doi.org/10.1016/j.chemolab.2015.05.006.

[20] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. Proc - Int Conf Pattern Recognit 2010:3121–4. https://doi.org/10.1109/ICPR.2010.764.

[21] MathWorks. Chapter 23: Classification learner. Stat. Mach. Learn. Toolbox User's Guid., 2019, p. 1–163.

[22] Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B 1958;20:215–42. https://doi.org/10.1007/BF03180993.

[23] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: John Wiley & Sons; 2001.

[24] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1023/A:1022627411411.

[25] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. vol. 19. 1984. doi:10.1371/journal.pone.0015807.

[26] Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43:1947–58. https://doi.org/10.1021/ci034160g.

[27] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man, Cybern Part ASystems Humans 2010;40:185–97. https://doi.org/10.1109/TSMCA.2009.2029559.

[28] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1–10. https://doi.org/10.1186/s12916-014-0241-z.

[29] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.

[30] Chajon E, Bellec J, Castelli J, Corre R, Kerjouan M, Le Prise E, et al. Simultaneously modulated accelerated radiation therapy reduces severe esophageal toxicity in concomitant chemoradiotherapy of locally advanced non-small-cell lung cancer. Br J Radiol 2015;88:1–4. https://doi.org/10.1259/bjr.20150311.

[31] Manapov F, Sepe S, Niyazi M, Belka C, Friedel G, Budach W. Dose-volumetric parameters and prediction of severe acute esophagitis in patients with locally-advanced non small-cell lung cancer treated with neoadjuvant concurrent hyperfractionated-accelerated chemoradiotherapy. Radiat Oncol 2013;8:1–4.

[32] Ahn S-J, Kahn D, Zhou S, Yu X, Hollis D, Shafman TD, et al. Dosimetric and clinical predictors for radiation-induced esophageal injury. Int J Radiat Oncol Biol Phys 2005;61:335–47. https://doi.org/10.1016/j.ijrobp.2004.06.014.

[33] Chapet O, Kong F, Lee JS, Hayman JA, Ten RK. Normal tissue complication probability modeling for acute esophagitis in patients treated with conformal radiation therapy for non-small cell lung cancer. Radiother Oncol 2005;77:176–81. https://doi.org/10.1016/j.radonc.2005.10.001.

[34] Belderbos J, Heemsbergen W, Hoogeman M, Pengel K, Rossi M, Lebesque J. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. Radiother Oncol 2005;75:157–64. https://doi.org/10.1016/j.radonc.2005.03.021.

[35] Huang EX, Robinson CG, Molotievschi A, Bradley JD, Deasy JO, Oh JH. Independent test of a model to predict severe acute esophagitis. Adv Radiat Oncol 2017;2:37–43. https://doi.org/10.1016/j.adro.2016.11.003.

[36] Thor M, Deasy J, Iyer A, Bendau E, Fontanella A, Apte A, et al. Toward personalized dose-prescription in locally advanced non-small cell lung cancer: validation of published normal tissue complication probability models. Radiother Oncol 2019;138:45–51. https://doi.org/10.1016/j.radonc.2019.05.011.

[37] Wada K, Kishi N, Kanayama N, Hirata T, Ueda Y, Kawaguchi Y, et al. Predictors of acute radiation esophagitis in non-small cell lung cancer patients treated with accelerated hyperfractionated chemoradiotherapy. Anticancer Res 2019;39:491–7. https://doi.org/10.21873/anticanres.13139.

[38] Rwigema JCM, Verma V, Lin L, Berman AT, Levin WP, Evans TL, et al. Prospective study of proton-beam radiation therapy for limited-stage small cell lung cancer. Cancer 2017;123:4244–51. https://doi.org/10.1002/cncr.30870.

[39] Hawkins PG, Boonstra PS, Hobson ST, Hayman JA, Ten Haken RK, Matuszak MM, et al. Prediction of radiation esophagitis in non-small cell lung cancer using clinical factors, dosimetric parameters, and pretreatment cytokine levels. Transl Oncol 2018;11:102–8. https://doi.org/10.1016/j.tranon.2017.11.005.

[40] Giaddui T, Chen W, Yu J, Lin L, Ii CBS, Yuan L, et al. Establishing the feasibility of the dosimetric compliance criteria of RTOG 1308: phase III randomized trial comparing overall survival after photon versus proton radiochemotherapy for inoperable stage II-IIIB NSCLC. Radiat Oncol 2016;11:1–7. https://doi.org/10.1186/s13014-016-0640-8.

[41] Luna JM, Chao HH, Diffenderfer ES, Valdes G, Chinniah C, Ma G, et al. Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning. Radiother Oncol 2019;133:106–12. https://doi.org/10.1016/j.radonc.2019.01.003.

[42] Chao HH, Valdes G, Luna JM, Heskel M, Berman AT, Solberg TD, et al. Exploratory analysis using machine learning to predict for chest wall pain in patients with stage I non-small-cell lung cancer treated with stereotactic body radiation therapy. J Appl Clin Med Phys 2018;19:539–46. https://doi.org/10.1002/acm2.12415.

[43] Bahn E, Alber M. On the limitations of the area under the ROC curve for NTCP modelling. Radiother Oncol 2020;144:148–51. https://doi.org/10.1016/j.radonc.2019.11.018.