Improved predictive modeling of radiation pneumonitis in lung cancer patients using machine learning techniques

Jung Hun Oh¹, Aditya Apte¹, Clifford G. Robinson², Jeffrey Bradley², and Joseph O. Deasy¹
¹Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, NY
²Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO

Introduction
Radiation pneumonitis (RP) is a potentially fatal side effect arising in lung cancer patients who receive radiotherapy as part of their treatment. For the modeling of RP outcomes data, several predictive models based on traditional statistical methods and machine learning techniques have been reported. However, no guidance to variation in performance has been provided to date.

Materials and Methods
1. Dataset
This dataset consists of 123 NSCLC patients who received 3D conformal radiation therapy with a median prescription dose of 66 Gy (46–74 Gy) as part of their treatment. With observations throughout a median follow-up period of 17 months for the endpoint of RP, the patients were divided into an RP group (n = 59; grade ≥ 2) and a control group (n = 64). Using our in-house software (DREES: dose-response explorer system)¹, lung and heart dose-volume variables were extracted.

2. Machine learning methods
2.1. Principle component analysis (PCA)
Principle component analysis (PCA) is a useful statistical method for dimension reduction. A key idea of PCA is to transform the original data space into a lower dimensional space with a minimum loss of information. As a result, a principle component (PC) score is expressed as a linear combination of the original variables, which captures the maximum variance for the data.

2.2. Logistic regression
Logistic regression is a widely used tool for modeling binary outcomes. The logistic regression model is defined as

\[ Y_i = \frac{\exp(x_i)}{1+\exp(x_i)} \]

where \( Y_i \) indicates outcomes for the \( i^{th} \) patient and \( x_i \) is a linear combination of the variables including a constant term, given by

\[ x_i = \beta_0 + \sum_{j=1}^{s} \beta_j x_{ij}. \]

2.3. Supervised PCA
Recently, Bair and his colleagues designed supervised PCA². In this approach, univariate regression coefficient for each variable against outcome is calculated. Then, variables whose coefficients exceed a predefined threshold \( \theta \) are used to perform PCA. As a result, the first \( m \) PCs are used to build a predictive model of the outcome.

2.4 Supervised PCA-based logistic regression
Our proposed method integrates the logistic regression with supervised PCA. That is, using variables that are selected by the evaluation of univariate logistic regression, PCA is performed. Then, with the first \( m \) PCs, predictive models are sought using multivariate logistic regression in conjunction with bootstrap statistical resampling.
Results

1. Univariate analysis

<table>
<thead>
<tr>
<th>Heart variables</th>
<th>Lung variables</th>
<th>Clinical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td>Rs</td>
<td>p</td>
</tr>
<tr>
<td>1 V40</td>
<td>0.25</td>
<td>0.0061</td>
</tr>
<tr>
<td>2 V45</td>
<td>0.24</td>
<td>0.0078</td>
</tr>
<tr>
<td>3 V5</td>
<td>0.23</td>
<td>0.0106</td>
</tr>
<tr>
<td>4 V35</td>
<td>0.23</td>
<td>0.0123</td>
</tr>
<tr>
<td>5 V10</td>
<td>0.22</td>
<td>0.0127</td>
</tr>
<tr>
<td>6 V50</td>
<td>0.22</td>
<td>0.0129</td>
</tr>
<tr>
<td>7 MOH55</td>
<td>0.22</td>
<td>0.0158</td>
</tr>
<tr>
<td>8 MOH60</td>
<td>0.22</td>
<td>0.0160</td>
</tr>
<tr>
<td>9 MOH50</td>
<td>0.22</td>
<td>0.0165</td>
</tr>
<tr>
<td>10 MOH70</td>
<td>0.22</td>
<td>0.0165</td>
</tr>
</tbody>
</table>

Vx (volume getting at least x Gy), Dx (minimum dose to the hottest x% volume), MOHx (mean dose to the hottest x% volume)

Table 1 shows the top 10 variables in each variable category including lung, heart, and clinical variables according to Spearman correlation coefficient (Rs) obtained using univariate logistic regression. Using a ‘Smoking’ variable, the best Rs was obtained (Rs=0.2742, p=0.0021). Overall heart variables had better Rs than lung variables. Interestingly, Vx and MOHx in heart variables were highly ranked, whereas overall Dx variables had high Rs values in lung variables.

2. Multivariate analysis

First, Rs for each variable was calculated using univariate logistic regression. The variables were sorted according to absolute Rs. Then, the proposed model was evaluated choosing the top 10, 15,...,30 variables. This procedure was iterated for lung, heart, and combined variables. Taking these variables, supervised PCA generated the same number of PCs. Using these PCs, bootstrap statistical resampling was performed to find the best model. The performance of the proposed method was compared with multivariate logistic regression alone using the top 10, 15,...,30 variables with original values. We limited the maximum number of predictors to 5 in the model. Integrating all variables, SPCA-based logistic regression model obtained the best performance with Rs=0.44 (p < 0.0001) while logistic model had Rs=0.37 (p < 0.0001). Interestingly, when clinical variables alone were used, logistic model showed better performance with Rs=0.34 (p < 0.0001) than SPCA-based logistic regression model with Rs=0.32 (p < 0.0001).

Conclusions

In this study, we demonstrated an application of machine-learning techniques for predicting RP in lung cancer patients. Incorporating multivariate logistic regression with a recently introduced machine learning method allowed us to produce a better predictive RP model that should be further tested for potential clinical use.

Reference