Published in final edited form as:

Med Phys. 2007 October ; 34(10): 3808–3814.

Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis

Shifeng Chen^{a)}, Sumin Zhou, Fang-Fang Yin, Lawrence B. Marks, and Shiva K. Das

Department of Radiation Oncology, Duke University Medical Center, Durham, North Carolina 27710

Abstract

The purpose of this study is to build and test a support vector machine (SVM) model to predict for the occurrence of lung radiation-induced Grade 2+ pneumonitis. SVM is a sophisticated statistical technique capable of separating the two categories of patients (with/without pneumonitis) using a boundary defined by a complex hypersurface. Despite the complexity, the SVM boundary is only minimally influenced by outliers that are difficult to separate. By contrast, the simple hyperplane boundary computed by the more commonly used and related linear discriminant analysis method is heavily influenced by outliers. Two SVM models were built using data from 219 patients with lung cancer treated using radiotherapy (34 diagnosed with pneumonitis). One model (SVM_{all}) selected input features from all dose and non-dose factors. For comparison, the other model (SVM_{dose}) selected input features only from lung dose-volume factors. Model predictive ability was evaluated using ten-fold cross-validation and receiver operating characteristics (ROC) analysis. For the model SVM_{all}, the area under the cross-validated ROC curve was 0.76 (sensitivity/specificity = 74%/75%). Compared to the corresponding SVM_{dose} area of 0.71 (sensitivity/specificity = 68%/68%), the predictive ability of SVM_{all} was improved, indicating that non-dose features are important contributors to separating patients with and without pneumonitis. Among the input features selected by model SVM_{all}, the two with highest importance for predicting lung pneumonitis were: (a) generalized equivalent uniform doses close to the mean lung dose, and (b) chemotherapy prior to radiotherapy. The model SVM_{all} is publicly available via internet access.

Keywords

support vector machines; radiation pneumonitis; modeling; prediction

I. INTRODUCTION

Lung radiation-induced pneumonitis is one of the major dose-limiting toxicities associated with thoracic radiotherapy (RT). To obtain the optimal balance between dose coverage to the target volume and minimization of the risk of radiation pneumonitis, it is important to understand the relationship between factors such as radiation dose-volume metrics and the incidence of radiation pneumonitis.

Several studies have suggested that the incidence of radiation pneumonitis depends on dose-volume factors (e.g., V_{20} (lung volume receiving dose above 20 Gy),^{1–9} mean lung dose,^{1,4,5,8,10–13} $V_{30}^{2,7,10}$ $V_{15}^{2,7}$ V_{40}^{7} and V_{50}^{7}), as well as non-dose factors (e.g., tumor

location,^{5,14} age,^{3,15} chemotherapy schedule,^{13,15} gender¹⁶). Most of these studies identify single/multiple factors (features) from univariate/multivariate analysis, but do not consider how these features may be combined into a predictive model. For example, in Lind *et al.*,³ the univariately correlated features are not, by themselves, strong predictors of radiation pneumonitis.³ It is possible that appropriately combining weakly correlated features into a model may yield much greater predictive accuracy.¹⁷ The aim of this work is to develop and test such a model.

We herein use the support vector machine (SVM)^{18,19} technique to assess predictors of radiation pneumonitis. SVM is a discriminative machine learning technique, based on Vapnik's structural risk minimization theory,¹⁸ which shares some similarities with linear discriminant analysis (LDA).²⁰ Like LDA, SVM uses a boundary to separate data points into two categories. Unlike LDA, which only uses a hyperplane as boundary, SVM is capable of complex hypersurfaces via a kernel function.²¹ Thus, SVM is more capable of segregating "clusters" of points sharing the same outcome (clusters in the space of the inputs) by using a closed hypersurface. SVM, unlike LDA, also tolerates some points on the wrong side of the boundary. By discarding the effect of these points, SVM reduces the possibility of a few outliers lying on the wrong side from casting an undue influence on the shape and location of the boundary. This feature improves model robustness and generalization. SVM has been successfully applied to problems of text categorization^{22,23} and face detection.²⁴

In this paper, we present and test an SVM model built using a novel feature selection algorithm. Model testing employed receiver operating characteristic (ROC) analysis^{3,25,26} and ten-fold cross-validation²¹ techniques. The importance of each SVM selected input feature was evaluated. This model is publicly available via the internet (http://www.radonc.duke.edu/modules/div_medphys/index.php?id=25).

II. MATERIALS AND METHODS

II.A. Input data

The study included 235 patients with lung cancer who received three-dimensional conformal radiotherapy at Duke University Medical Center on an Institutional Review Board approved protocol. Radiation-induced symptomatic pneumonitis was diagnosed and graded at follow-up (typically at 1, 3, and then every 3–4 months post-radiotherapy). Pneumonitis was graded from 0 to 4, as follows: Grade 0: no increase in symptoms; Grade 1: symptoms not requiring initiation or increase in steroids and/or oxygen; Grade 2: symptoms requiring initiation or increase in steroids; Grade 3: symptoms requiring oxygen; Grade 4: symptoms requiring assisted ventilation or causing death. Among these patients, 34 were diagnosed with Grade 2+ pneumonitis and 16 were classified as "hard-to-score,"²⁷ i.e., uncertain diagnosis of Grade 2+ pneumonitis. The "hard-to-score" patients were excluded in this analysis.

A total of 93 parameters were collected for each patient, consisting of dose and relevant non-dose factors. The dose factors included mean heart dose, the lung dose-volume histogram (DVH) (percentage of lung volume above dose ranging from 6 to 60 Gy in increments of 2 Gy), and 37 lung generalized equivalent uniform doses (EUD).²⁸ EUD was calculated as

$$\text{EUD} = \left(\frac{\sum_i V_i D_i^a}{\sum_i V_i} \right)^{1/a}, \quad (1)$$

where V_i is the lung volume receiving dose D_i , and a is a parameter ranging from 0.4 to 4 in increments of 0.1. Note that for $a=1$, EUD is equivalent to mean lung dose. Non-dose factors included race, age, gender, tumor stage, tumor location (central or peripheral; upper, middle or lower lobe; right or left lung), chemotherapy schedule (none, pre-RT, concurrent, pre-RT and concurrent, post-RT, or concurrent and post-RT), histology type (squamous cell, adenocarcinoma, non-small-cell, small-cell, large-cell, or other), surgery (yes or no), once or twice daily RT, pre-RT FEV₁ (forced expiratory volume in 1 s), FEV₁% (as percentage of predicted normal), pre-RT DLCO (carbon monoxide diffusion capacity in lung), and pre-RT DLCO% (as percentage of predicted normal). These factors were evaluated for selection as SVM input features, according to the procedure described later.

II.B. Ten-fold cross-validation (model testing)

The patient data were randomly split into 10 approximately equal-sized groups. Nine groups (training data) were used to train the SVM and the remaining group (cross-validation data) was used as a test to measure the performance of the SVM. Note that feature selection was based only on the training data. The procedure was repeated ten times, with each group, in turn, serving as the test set.

II.C. Support vector machines for classification

The general idea behind SVMs is to compute an optimal hypersurface (boundary) that maximizes the margin between data points in categories +1 (radiation pneumonitis) and -1 (no radiation pneumonitis). Figure 1 illustrates this concept using a simplified hyperplane. Each data point, x is a vector containing the list of selected features (features selected from the 93 factors available for each patient). In Fig. 1, the hyperplane is denoted by $x^T\beta + \beta_0 = 0$, where β is the normal vector to the hyperplane, and β_0 is an offset parameter ($|\beta_0|/\|\beta\|$ is the closest distance from the origin to the hyperplane). The two parallel margin edges are $x^T\beta + \beta_0 = \pm 1$, corresponding to a margin width of $2/\|\beta\|$. Physically, β may be interpreted as the importance weights assigned to the selected features and β_0 as the separation threshold of the weighted combination of features (separating cases with and without pneumonitis). Ideally, the space within the margin is devoid of data points. In reality, SVM maximizes the margin [distance between dotted lines in Fig. 1(a)], while permitting some transgression of points into the margin area or the wrong side of the boundary [Fig. 1(b)]. In Fig. 1(b), the slack variable $\xi'_i > 0$ indicates that the point i has transgressed into the margin, and $\xi'_i > 1/\|\beta\|$ indicates that it is on the wrong side of the separation hyperplane. The sum total of transgressions is restricted to be below a certain limit, thereby allowing small transgressions from several points or, alternately, large outlier transgressions from a few points.

Mathematically, the concept of SVM is described as

$$\max \frac{2}{\|\beta\|} \text{ subject to } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \\ \xi_i \geq 0, \sum \xi_i \leq \text{const}, \end{cases} \quad (2)$$

where $\xi_i = \|\beta\| \xi'_i$ [see Fig. 1(b)], (x_i, y_i) is the data point for patient i , with $y_i = \pm 1$ (+1: radiation pneumonitis; -1: no radiation pneumonitis) and vector x_i containing the features selected from the 93 available patient factors.

The solution of Eq. (2) is equivalent to solving a quadratic programming problem²¹ in which

$$\min_{\alpha_i} \left(\frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^N \alpha_i \right) \quad (3)$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^N \alpha_i y_i = 0$. Parameter C is user defined. It is determined as explained in the next subsection. The coefficients α_i are computed through the optimization, and the optimized coefficients are denoted by $\hat{\alpha}_i$.

The solution for β has the form

$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i y_i x_i, \quad (4)$$

where $\hat{\alpha}_i$ [solution of Eq. (3)] is nonzero only for data points that transgress the margin. The position vectors of these points are the “support vectors” that define the hyperplane normal. Points on the margin boundary are used to solve for β_i . Thus, the final SVM classifier is

$$\text{sign}(x^T \hat{\beta} + \beta_0). \quad (5)$$

The support vector machine shown in Fig. 1 has a linear boundary. However, for cases with nonlinear boundaries, the original low-dimensional input space can be translated to a higher-dimensional feature space via a basis function²¹ (see Fig. 2). The basis function does not explicitly appear in the calculation if $x_i^T x_j$ [Eq. (3)] is replaced by an appropriate kernel function $K(x_i, x_j)$.²¹ This mapping allows the SVM to accommodate nonlinear boundaries that achieve better training-class separation.²¹

There are several possible choices for the kernel function K .²¹ In this work, the radial basis function, a popular kernel function in SVM literature,^{18,19,21} was used:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2), \quad (6)$$

where σ is a user-defined parameter, determined as described in the next subsection. Two other choices for $K(x_i, x_j)$ found in SVM literature²¹ are

$$d\text{th degree polynomial: } K(x_i, x_j) = (1 + x_i^T x_j)^d, \quad (7)$$

$$\text{sigmoid: } K(x_i, x_j) = \tanh(k_1 x_i^T x_j + k_2)^d. \quad (8)$$

The first function [Eq. (7)] can be problematic, since it is unbounded and can potentially lead to numerical instability. The second function [Eq. (8)] has two free parameters and hence is more likely to overfit the model, compared to the radial basis kernel function with one free parameter. Therefore, the radial basis kernel function was used in this work.

In summary, training an SVM is equivalent to solving a quadratic programming problem [Eq. (3)] with $x_i^T x_j$ replaced by the kernel function $K(x_i, x_j)$.

II.D. Parameter selection

Parameters C [Eq. (3)] and σ [Eq. (6)] were determined prior to cross-validation using grid search²⁹ and nine-fold evaluation within each training set. Of the nine groups constituting each training set (see Fig. 1), eight groups were used to train the SVM using a specific (C, σ) value, and one group (termed training-evaluation) was used to evaluate the SVM for prediction accuracy using ROC analysis. The procedure was repeated nine times for each training set, such that each group was used once for training-evaluation. The parameters (C, σ) were optimized in the grid space of $(\log_{10} C, \log_{10} \sigma^2)$ ²⁹ to maximize the area under the training-evaluation ROC curve. Once determined, the optimal (C, σ) values were applied to the entire training set for cross-validated testing.

II.E. Input feature selection

Each of the 93 patient variables is potentially an input feature for the SVM. Input features were selected using a unique algorithm that progressively built the SVM by sequentially adding/substituting input features.

Input features were selected using the nine-fold training-evaluation scheme described in the previous subsection. For each variable that was a potential input feature, the SVM was trained using eight of the nine training groups and then evaluated on the one remaining training-evaluation group. Each of the nine training groups served as the training-evaluation group, in turn. The variable was added as input feature if the area under the collective ROC curve of the nine training-evaluation groups increased. Similarly, an already selected input feature was replaced by another variable if the training-evaluation ROC area increased. SVM construction was stopped if no new variable was accepted as input feature, after all unselected variables were evaluated through addition/substitution.

In summary, the algorithm is as follows (“AUC” denotes the area under ROC curve for training-evaluation):

Operator₁: substitute one existing input feature with one unselected variable

Operator₂: add one input feature

Step 1: do for each Operator_k, $k=1,2$

Step 2: do for each variable $i=1, \dots, 93$ ($i \neq$ the variable selected as input features)

1. Trial new SVM: $SVM^* = \text{Operator}_k(SVM)$

2. ROC evaluation;

$$\begin{aligned} AUC_{SVM^*} &> AUC_{SVM}, \\ SVM &\leftarrow SVM^*, \text{ go to step 1;} \end{aligned}$$

3. If $k=2$, $i=93$, and SVM is not replaced, stop construction.

When there are less than three input features, Operator₁ (substitution) was skipped.

For the purpose of ten-fold cross-validation, the procedure above was repeated ten times, corresponding to each cross-validation group.

II.F. Model comparison

To evaluate the effect of non-dose input features on SVM prediction accuracy, two SVM models were built and tested. The first SVM model (SVM_{all}) had input features selected from all dose and non-dose variables, while the second SVM model (SVM_{dose}) had input features selected only from lung dose-volume histogram variables. The cross-validated ROC AUCs from the two models were used to compare generalization capability (larger AUC implies a more accurate model).

II.G. Evaluation of importance of input features

The importance of each input feature was evaluated by excluding it from the model, one feature at a time. For each exclusion, parameters C and σ were estimated and the SVM was trained as explained in the previous subsections. With the exclusion of each input feature, the decrement in cross-validated AUC was used to rank the importance of the excluded feature (larger decrement denotes a more important feature).

III. RESULTS AND DISCUSSION

III.A. Model building

The SVM algorithm was programmed in-house, using MATLAB (Mathworks, Natick, MA). For the purpose of tenfold cross-validation, 10 SVMs were built, with each SVM used to evaluate one of the test groups. Thus, the resulting model SVM_{all} or SVM_{dose} is an ensemble of ten component SVMs. Note that, since the SVMs were built and trained with slightly different training data (2/9 of the training data are different between any two SVMs), they understandably have different input features and values of the parameters C and σ .

The selected input features are listed in Table I for models SVM_{all} and SVM_{dose} . This table lists, in brackets, the number of component SVMs that selected a specific input feature. For model SVM_{all} , each component SVM selected four input features. Note that the generalized equivalent uniform doses $EUD\ a=1.2$, 1.3 , and 1.4 are highly correlated to each other. However, they were selected by different component SVMs, i.e., no two of these correlated features were selected by the same component SVM. $EUD\ a=1.3$ was selected by seven component SVMs, $EUD\ a=1.4$ by two component SVMs, and $EUD\ a=1.2$ by 1 component SVM. $EUD\ a=1.2$, 1.3 , and 1.4 are highly correlated to $EUD\ a=1$ (mean lung dose), which frequently appears as a strong predictor of radiation pneumonitis in literature.^{1,4,5,8,10-13} Thus, mean lung dose appears to be a stronger predictive parameter than any lung DVH V_x metric (V_x metrics are frequently reported in literature¹⁻¹⁰). A possible explanation for the stronger predictive potential of generalized EUD metrics is that they contain information on the DVH shape (by volume-weighted combination of dose), as opposed to V_x metrics that only select a single point on the DVH curve.

All ten component SVMs chose chemotherapy-prior-to-RT as an input feature that was predictive for radiation pneumonitis. This factor also appears in literature as associated with the occurrence of pneumonitis.^{13,15,30} McDonald *et al.*³⁰ report that, while some chemotherapeutic drugs can induce lung injury such as pneumonitis, chemotherapeutic drugs can also enhance radiation-induced lung injury. Other input features, ranked in order from most-to-least commonly selected, were: Tumor location (central or peripheral) selected by nine component SVMs, gender (male or female) by eight component SVMs, histology (adenocarcinoma or not) by two component SVMs, and histology (small cell or not) by one component SVM. Peripheral tumor location, female sex, and small cell/adenocarcinoma histology were associated with increased risk of pneumonitis. Female sex has been implicated in prior studies,¹⁶⁻³¹ whereas histology and peripheral tumor location have not

(inferior tumor location has previously been identified as correlated with pneumonitis¹⁴). However, it is not surprising that the SVM algorithm sometimes selects input features with weak univariate correlation, since multiple such variables could synergistically interact (when combined) to provide high correlation (“..., a variable that is completely useless by itself can provide a significant performance improvement when taken with others”¹⁷). Indeed, one of the benefits of this approach is the ability to consider such complex interactions.

For model SVM_{dose}, each component SVM selected two input features. Eight component SVMs selected EUD $a=1.3$, while highly correlated EUD $a=1.4$ and EUD $a=1.1$ were each selected by one component SVM each. Seven SVMs selected V_{50} and three SVMs selected the closely related V_{48} . The selection of only two input features from the lung DVH is likely because lung DVH variables tend to be highly correlated to each other.

III.B. Parameter selection

While the concept of a support vector machine is straight-forward, “incorrect” values of the parameters C [Eq. (3)] and σ [Eq. (6)] can lead to poor predictive accuracy. These parameters control the tradeoff between the model overfitting (the model is too complex—it fits the signal as well as the noise) and underfitting (the model is not complex enough to fit the signal). Thus, it is critical to accurately estimate their values for good generalization.

Parameter C controls the complexity of the hypersurface that separates the two categories. For small values of C , the separation hypersurface created by the SVM algorithm can be insufficiently complex, resulting in underfitting. As C increases, the SVM algorithm increases the complexity of the separation hypersurface to correctly classify greater numbers of data points. Thus, large values of C can lead to overfitting. The effect of parameter C is shown in Fig. 3. In Fig. 3, the AUC for training-evaluation is shown as a function of $\log_{10} C$, for fixed σ ($\log_{10} \sigma^2 = -1.5$). For $\log_{10} C < 4$, the SVM underfits the data. The overfitting condition is not shown here.

Parameter σ is the width of the radial basis function, which controls the number of support vectors. For smaller σ , the SVM uses more data points as support vectors, leading to overfitting. Conversely, for larger σ , the SVM has fewer support vectors, leading to underfitting. The behavior of σ is shown in Fig. 4 (parameter C was fixed at $\log_{10} C = 5$). The low values of AUC for small σ represent underfitting. Underfitting is reduced with increasing σ , up to $\log_{10} \sigma^2 = -2$. Overfitting is manifested as a sharp drop in the area under the training-evaluation ROC curve, beyond $\log_{10} \sigma^2 = -0.5$.

The optimal values of parameters C and σ (optimized using grid search²⁹ to maximize the training-evaluation AUC) are not a single point, but rather an area in the space of ($\log_{10} C$, $\log_{10} \sigma^2$). As seen in the examples of Fig. 4 and Fig 5, C is optimal for $\log_{10} C > 4$ when $\log_{10} \sigma^2 = -1.5$ (Fig. 3), and σ is optimal for $\log_{10} \sigma^2 = [-2, -0.5]$ when $\log_{10} C = 8$ (Fig. 4).

III.C. Modeled testing and comparison

The ROC analysis results for ten-fold cross-validated testing are shown in Fig. 5 for the SVM using dose and non-dose variables (SVM_{all}), and in Fig. 6 for SVM using only lung DVH variables (SVM_{dose}). For model SVM_{all}, the area under the ROC curve (AUC) was 0.76 (sensitivity=74%, specificity=75%), while for model SVM_{dose}, the AUC was 0.71 (sensitivity=68%, specificity=68%). The difference between these two areas suggests that the predictive ability of model SVM_{all} is better than that of model SVM_{dose} and that the addition of non-dose features can improve the generalization capability of the SVM model.

III.D. Evaluation of importance of input features

The importance of each input feature used in model SVM_{all} was evaluated and ranked, from highest to lowest, as follows: dose metrics closely related to mean lung dose (EUD $a=1.2$, 1.3, and 1.4), chemotherapy before radio-therapy (yes or no), tumor position (central or peripheral), gender (male or female), histology (adenocarcinoma or not), and histology (small-cell or not). The ROC analysis is summarized in Table II. The exclusion of dose metrics closely related to mean lung dose resulted in a large AUC drop, from 0.76 to 0.57. This suggests, in agreement with other studies^{1,4,5,8,10-13} that mean lung dose metrics are an important factor in predicting lung pneumonitis. The second most important feature is chemotherapy prior to radiotherapy (AUC decrement of 0.09). The use of chemotherapy, either prior to or concurrent with RT, has been suggested in other studies to increase the risk of pneumonitis.³⁰ Even though other input features record only small drops in the AUC, they nevertheless help to improve model generalization. It is understandable that histology (adenocarcinoma or not) and histology (small-cell or not) would have minimal impact on AUC reduction, since they were only selected by two component SVMs and one component SVM, respectively (see Table I).

III.E. Evaluation of the robustness of cross-validated results to patient assignment

To evaluate the robustness of the cross-validated results from model SVM_{all} to patient assignment (i.e., sensitivity of the results to randomization of patients into the ten different groups), the data were randomly split 100 times into ten groups. Thus, each time, the composition of patients within the ten groups changed. Each time, model SVM_{all} was trained and tested with ten-fold cross-validation. The cross-validated ROC areas from the 100 randomizations had mean=0.74 (range 0.71–0.77) and standard deviation=0.03. The small variance implies that the dataset is of adequate size and that model SVM_{all} is robust.

III.F. Model publication

The model SVM_{all} for prospective use is available for download from http://www.radonc.duke.edu/modules/div_medphys/index.php?id=25. The required input features are shown in the left column of Table I. The input file (example available on website) is required to include the entire lung DVH, chemotherapy prior to RT (yes or no), tumor position (central or peripheral), gender (male or female), histology (adenocarcinoma or not), and histology (small-cell or not). Missing variables are indicated as negative values in the input file. The program internally computes two of the input features from the lung DVH: EUDs with $a=1.2$, 1.3, and 1.4. The classification result is an average of outputs from the ten component SVMs comprising SVM_{all}. The model outputs are two sets of metrics: a discriminant value that is a measure of the extent of injury (>0 indicates predicted pneumonitis, <0 indicates no predicted pneumonitis), and the number of patients in the Duke training database with higher discriminant than the prospectively tested patient. The latter value ranks the prospectively evaluated patient in the context of the Duke population.

IV. CONCLUSIONS

In this work, the support vector machine (SVM) algorithm was investigated to predict lung radiation-induced pneumonitis. Results indicate that the SVM model is a powerful, yet robust, predictor. The SVM model constructed with dose and non-dose input features yielded a ten-fold cross validated ROC area of 0.76 with sensitivity and specificity of 74% and 75%, respectively. Among the selected input features, dose metrics closely related to mean lung dose were most influential. The SVM model constructed in this work is available for public use via internet access.

Acknowledgments

This work was supported by Grant Nos. NIH R01 CA 115748 and NIH R01 CA69579.

References

1. Kong F-M, et al. Final toxicity results of a radiation-dose escalation study in patients with non-small-cell lung cancer (NSCLC): predictors for radiation pneumonitis and fibrosis. *Int. J. Radiat. Oncol., Biol., Phys* 2006;65:1075–1086. [PubMed: 16647222]
2. Tsujino K, et al. Radiation pneumonitis following concurrent accelerated hyperfractionated radiotherapy and chemotherapy for limited-stage small-cell lung cancer: Dose-volume histogram analysis and comparison with conventional chemoradiation. *Int. J. Radiat. Oncol., Biol., Phys* 2006;64:1100–1105. [PubMed: 16373082]
3. Lind PA, et al. ROC curves and evaluation of radiation-induced pulmonary toxicity in breast cancer. *Int. J. Radiat. Oncol., Biol., Phys* 2006;64:765–770. [PubMed: 16257129]
4. Chang DT, et al. The impact of heterogeneity correction on dosimetric parameters that predict for radiation pneumonitis. *Int. J. Radiat. Oncol., Biol., Phys* 2006;65:125–131. [PubMed: 16427214]
5. Graham MV, et al. Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC). *Int. J. Radiat. Oncol., Biol., Phys* 1999;45:323–329. [PubMed: 10487552]
6. Tsujino K, et al. Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. *Int. J. Radiat. Oncol., Biol., Phys* 2003;55:110–115. [PubMed: 12504042]
7. Rancati T, et al. Factors predicting radiation pneumonitis in lung cancer patients: a retrospective study. *Radiother. Oncol* 2003;67:275–283. [PubMed: 12865175]
8. Jenkins P, et al. Radiation pneumonitis following treatment of non-small-cell lung cancer with continuous hyperfractionated accelerated radiotherapy (CHART). *Int. J. Radiat. Oncol., Biol., Phys* 2003;56:360–366. [PubMed: 12738310]
9. Moiseenko V, et al. Dose-volume analysis of lung complications in the radiation treatment of malignant thymoma: a retrospective review. *Radiother. Oncol* 2003;67:265–274. [PubMed: 12865174]
10. Hernando ML, et al. Radiation-induced pulmonary toxicity: A dose-volume histogram analysis in 201 patients with lung cancer. *Int. J. Radiat. Oncol., Biol., Phys* 2001;51:650–659. [PubMed: 11597805]
11. Martel MK, Tenhaken RK, Hazuka MB, et al. Dose-volume histogram and 3-D treatment planning evaluation of patients with pneumonitis. *Int. J. Radiat. Oncol., Biol., Phys* 1994;28:575–581. [PubMed: 8113100]
12. Kwa SLS, et al. Radiation pneumonitis as a function of mean lung dose: An analysis of pooled data of 540 patients. *Int. J. Radiat. Oncol., Biol., Phys* 1998;42:1–9. [PubMed: 9747813]
13. Theuvs JCM, et al. Prediction of over-all pulmonary function loss in relation to the 3-D dose distribution for patients with breast cancer and malignant lymphoma. *Radiother. Oncol* 1998;49:233–243. [PubMed: 10075256]
14. Hope AJ, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int. J. Radiat. Oncol., Biol., Phys* 2006;65:112–124. [PubMed: 16618575]
15. Theuvs JCM, et al. Dose-effect relations for early local pulmonary injury after irradiation for malignant lymphoma and breast cancer. *Radiother. Oncol* 1998;48:33–43. [PubMed: 9756170]
16. Robnett TJ, Machtay M, Vines EF, et al. Factors predicting severe radiation pneumonitis in patients receiving definitive chemoradiation for lung cancer. *Int. J. Radiat. Oncol., Biol., Phys* 2000;48:89–94. [PubMed: 10924976]
17. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res* 2003;3:1157–1182.
18. Vapnik, V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag; 1995.
19. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov* 1998;2(2):121–167.

20. Rencher, AC. *Methods of Multivariate Analysis*. 2nd ed.. New York: Wiley; 2002.
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag; 2002.
22. Kindermann J, Leopold E. Text categorization with support vector machines. *Mach. Learn* 2002;46:423–444.
23. Joachims, T. Presented at the 10th European Conference on Machine Learning; Germany: Chemnitz; 1998.
24. Freund, R.; Osuna, E.; Girosi, F. Presented at the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97) San Juan; Puerto Rico: 1997.
25. Das SK, et al. Predicting radiotherapy-induced cardiac perfusion defects. *Med. Phys* 2005;32:19–27. [PubMed: 15719950]
26. Swets, J.; Pickett, R. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press; 1982.
27. Kocak Z, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. *Int. J. Radiat. Oncol., Biol., Phys* 2005;62:635–638. [PubMed: 15936538]
28. Niemierko A. A generalized concept of equivalent uniform dose (EUD). *Med. Phys* 1999;26:1100–1100.
29. Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput* 2003;15:1667–1689. [PubMed: 12816571]
30. McDonald S, et al. Injury to the lung from cancer-therapy—clinical syndromes, measurable end-points, and potential scoring systems. *Int. J. Radiat. Oncol., Biol., Phys* 1995;31:1187–1203. [PubMed: 7713782]
31. Das SK, et al. Predicting lung radiotherapy-induced pneumonitis using a model that combines parametric Lyman probit with non-parametric decision trees. *Int. J. Radiat. Oncol., Biol., Phys* 2007;68:1212–1221. [PubMed: 17637394]

FIG. 1.

Support vector machines for (a) separable and (b) non-separable cases. The solid line is the optimal hyperplane separating the two categories. The dotted lines are margin edges with maximal width $2/\|\beta\|$. Points indicated by arrows are located within the margin or on the wrong side of the hyperplane (distance ξ_i from the margin edge). These points determine the separation hyperplane and are hence termed "support vectors."

FIG. 2. Data points are mapped from a low-dimensional input space to a high-dimensional feature space, thereby mapping the nonlinear boundary into a linear boundary. The mapping is achieved via a kernel function.

FIG. 3.
Area under the ROC curve (AUC) for training-evaluation data versus parameter $\log_{10} C$.
Parameter σ was fixed ($\log_{10} \sigma^2 = -1.5$).

FIG. 4.
Area under the ROC curve (AUC) for training-validation data versus parameter $\log_{10} \sigma^2$.
Parameter C was fixed ($\log_{10} C=8$).

FIG. 5. Cross-validated ROC curve for model SVM_{all} . The area under the curve, sensitivity, and specificity are 0.76, 74%, and 75%, respectively.

FIG. 6. Cross-validated ROC curve for model SVM_{dose} . The area under the curve, sensitivity, and specificity are 0.71, 68%, and 68%, respectively.

Table I

The input features of model SVM_{all} and SVM_{dose} , respectively. Shown in brackets is the number of component SVMs that selected a specific feature (ten component SVMs were built for cross-validation).

SVM_{all}	SVM_{dose}
EUD $a=1.3$ (7)	EUD $a=1.3$ (8)
EUD $a=1.2$ (1)	EUD $a=1.4$ (1)
EUD $a=1.4$ (2)	EUD $a=1.1$ (1)
Chemotherapy prior to RT (yes or no) (10)	Lung volume receiving dose >48 Gy (3)
Tumor location (central or peripheral) (9)	Lung volume receiving dose >50 Gy (7)
Gender (male or female) (8)	
Histology (adenocarcinoma or not) (2)	
Histology (small-cell or not) (1)	

Table II

Evaluation of importance of input features by single exclusion.

Input feature excluded	AUC decrement
EUD $a=1.1, 1.2, 1.3$	0.19
Chemotherapy prior to radiotherapy (yes or no)	0.09
Tumor position (central or peripheral)	0.02
Gender (male or female)	0.02
Histology (adenocarcinoma or not)	0.01
Histology (small-cell or not)	$<10^{-2}$