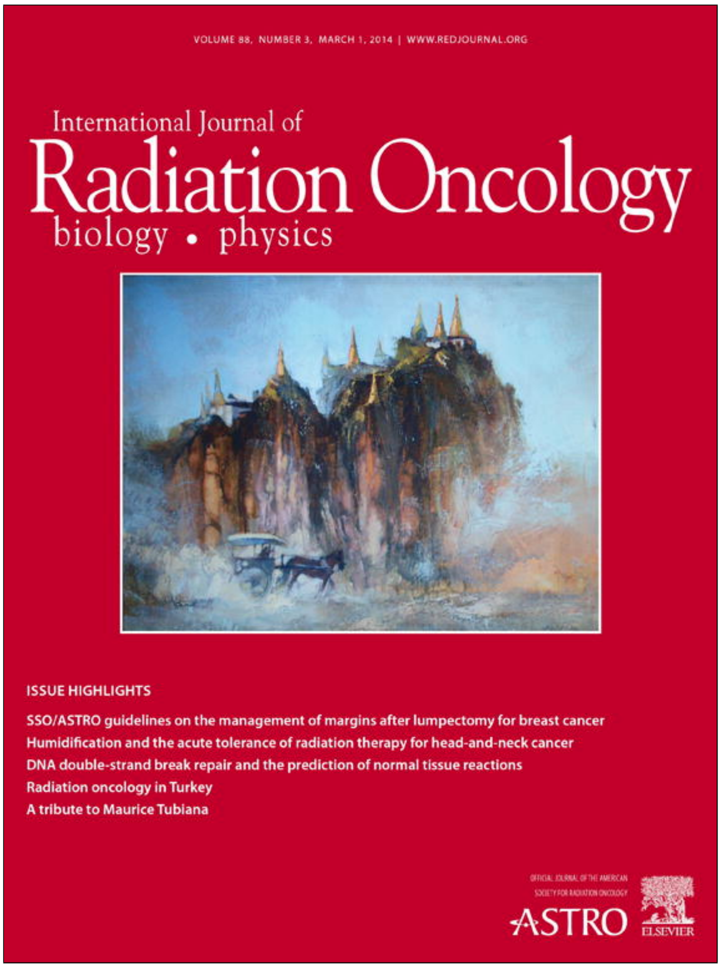
VOLUME 88, NUMBER 3, MARCH 1, 2014 | WWW.REDJOURNAL.ORG
International Journal of
Radiation Oncology
biology • physics
ISSUE HIGHLIGHTS
SSO/ASTRO guidelines on the management of margins after lumpectomy for breast cancer
Humidification and the acute tolerance of radiation therapy for head-and-neck cancer
DNA double-strand break repair and the prediction of normal tissue reactions
Radiation oncology in Turkey
A tribute to Maurice Tubiana
OFFICIAL JOURNAL OF THE AMERICAN SOCIETY FOR RADIATION ONCOLOGY
ASTRO
ELSEVIER

International Journal of
Radiation Oncology
biology ● physics

www.redjournal.org

Physics Contribution

# Support Vector Machine-Based Prediction of Local Tumor Control After Stereotactic Body Radiation Therapy for Early-Stage Non-Small Cell Lung Cancer

Rainer J. Klement, PhD,[*,†] Michael Allgäuer, MD,[‡] Steffen Appold, MD,[§] Karin Dieckmann, MD,[||] Iris Ernst, MD,[¶] Ute Ganswindt, MD,[#] Richard Holy, MD,[**] Ursula Nestle, MD,[††] Meinhard Nevinny-Stickel, MD,[‡‡] Sabine Semrau, MD,[§§] Florian Sterzing, MD,[||||] Andrea Wittig, MD,[¶¶] Nicolaus Andratschke, MD,[##] and Matthias Guckenberger, MD[*]

*[*]Department of Radiation Oncology, University of Würzburg, Germany, [†]Department of Radiotherapy and Radiation Oncology, Leopoldina Hospital, Schweinfurt, Germany, [‡]Department of Radiotherapy, Barmherzige Brüder Regensburg, Regensburg, [§]Department of Radiation Oncology, Technische Universität Dresden, Germany, [||]Department of Radiotherapy, Medical University of Vienna, Austria, [¶]Department of Radiotherapy, University Hospital Münster, Germany, [#]Department of Radiation Oncology, Ludwigs-Maximilians-University Munich, München, Germany, [**]Department of Radiation Oncology, RWTH Aachen University, Aachen, Germany, [††]Department of Radiation Oncology, University Hospital Freiburg, Freiburg i Br, Germany, [‡‡]Department of Therapeutic Radiology and Oncology, Innsbruck Medical University, Austria, [§§]Department of Radiation Oncology, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany, [||||]Department of Radiation Oncology, University Hospital Heidelberg, Germany, [¶¶]Department of Radiotherapy and Radiation Oncology, Philipps-University Marburg, Germany; and [##]Department of Radiation Oncology, Technische Universität München, Germany*

Received Jul 23, 2013, and in revised form Nov 8, 2013. Accepted for publication Nov 13, 2013.

**Summary**

We show that a support vector machine (SVM) classifier outperforms a multivariate logistic model in predicting tumor control probability after stereotactic body radiation therapy for early stage

**Background:** Several prognostic factors for local tumor control probability (TCP) after stereotactic body radiation therapy (SBRT) for early stage non-small cell lung cancer (NSCLC) have been described, but no attempts have been undertaken to explore whether a nonlinear combination of potential factors might synergistically improve the prediction of local control.
**Methods and Materials:** We investigated a support vector machine (SVM) for predicting TCP in a cohort of 399 patients treated at 13 German and Austrian institutions. Among 7 potential input features for the SVM we selected those most important on the basis of forward feature selection, thereby evaluating classifier performance by using 10-fold cross-validation and computing the area under the ROC curve (AUC). The final SVM classifier was built by repeating the feature selection 10 times with different splitting of the data for cross-validation and finally choosing

Reprint requests to: Dr Rainer J. Klement, PhD, Department of Radiotherapy and Radiation Oncology, Leopoldina Hospital Schweinfurt, Gustav-Adolf-Straße 8, 97422 Schweinfurt, Germany. Tel: (+49) 9721-7206424; E-mail: rainer_klement@gmx.de

Conflict of interest: none.

Supplementary material for this article can be found at www.redjournal.org.

*Acknowledgments*—We thank 3 anonymous reviewers for their valuable comments that significantly improved a previous version of our manuscript. We also would like to thank Dr. Colin E. Champ for proofreading the manuscript.

Int J Radiation Oncol Biol Phys, Vol. 88, No. 3, pp. 732–738, 2014
0360-3016/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijrobp.2013.11.216

non-small cell lung cancer in a cohort of 399 patients. Sensitivity and specificity of the SVM were 67.0% ± 0.5% and 78.7% ± 0.3%, respectively. These results suggest that machine learning techniques can be applied successfully to improve tumor control probability predictions.

only those features that were selected at least 5 out of 10 times. It was compared with a multivariate logistic model that was built by forward feature selection.

**Results:** Local failure occurred in 12% of patients. Biologically effective dose (BED) at the isocenter ($BED_{ISO}$) was the strongest predictor of TCP in the logistic model and also the most frequently selected input feature for the SVM. A bivariate logistic function of $BED_{ISO}$ and the pulmonary function indicator forced expiratory volume in 1 second (FEV1) yielded the best description of the data but resulted in a significantly smaller AUC than the final SVM classifier with the input features $BED_{ISO}$, age, baseline Karnofsky index, and FEV1 (0.696 ± 0.040 vs 0.789 ± 0.001, $P<.03$). The final SVM resulted in sensitivity and specificity of 67.0% ± 0.5% and 78.7% ± 0.3%, respectively.

**Conclusions:** These results confirm that machine learning techniques like SVMs can be successfully applied to predict treatment outcome after SBRT. Improvements over traditional TCP modeling are expected through a nonlinear combination of multiple features, eventually helping in the task of personalized treatment planning. © 2014 Elsevier Inc.

## Introduction

Stereotactic body radiation therapy (SBRT) is the treatment of choice for inoperable patients with early-stage non-small cell lung cancer (NSCLC). Clinical studies have shown that SBRT can result in excellent local control rates exceeding 90% with concurrently low toxicity rates, but efforts to accurately model the probability of tumor control (TCP) are still ongoing.

A dose−response relationship between the biologically effective dose (BED) and TCP in NSCLC is well established (1-4). BED is defined as the total dose that is needed to achieve the same biological effect in a tumor or organ as the treatment schedule under consideration if infinitesimally small doses were to be applied in an infinitely large number of fractions. Besides dose, it has been shown that nondosimetric factors such as tumor volume (5), glucose metabolic rate (6), tumor hypoxia (7), and oncogene activation (8) may play a fundamental role in determining tumor control after radiation in NSCLC. This implies that the full wealth of dosimetric, clinical, imaging, and molecular data now available for individual patients should be used together to obtain the highest possible accuracy of outcome predictions and aid in clinical decision support (9). Currently, however, most outcome predictions for NSCLC patients treated with SBRT rely on simple cutoffs (10, 11) or on fitting a logistic TCP function, with only a few attempts to incorporate other features besides dose to improve TCP predictions (5, 7). Such predictive models have successfully been used to identify dose−response relationships and to establish critical dose thresholds for achieving high tumor control rates >90% (12, 13). However, they also leave the investigator with the difficult task of identifying and modeling the interaction between variables that determines the outcome, whereas their mathematical framework often lacks the flexibility to realistically model such interactions. In this regard, machine learning techniques may be a better alternative because they allow for combining several features to build adaptive models based on the information contained in the data and thus do not depend on assumed mathematical relationships between dose and response (8, 14).

In machine learning, a classifier is trained on a set of data with known class labels so that it "learns" the distribution of the different classes in a multidimensional feature space (15). Machine learning algorithms are useful tools for data mining approaches (ie, the systematic investigation of all available data with the goal of discovering new patterns and new predictive variables that could lead to better prediction accuracy and insights into causative factors). For example, Chen et al (16) showed that by combining dosimetric and patient-specific features in a support vector machine (SVM), the prediction of severe radiation-induced pneumonitis in NSCLC patients could be improved compared with using dosimetric quantities alone. Similarly, Naqa et al (14) demonstrated that SVMs performed better than both multivariate logistic regression and mechanistic radiobiological models in predicting TCP for a set of 56 NSCLC patients treated with 3-dimensional conformal radiation therapy, particularly for those at high risk for local failure. Using a Bayesian network approach, Oh et al (17) revealed the usefulness of inflammatory and hypoxia biomarkers in addition to treatment plan-related variables for improving local control predictions after radiation therapy in advanced NSCLC patients. Finally, SVM-based integration of multidimensional gene expression profiling data has led to improved outcome predictions in various cancers, including breast (18) and nasopharyngeal carcinoma (19).

In this work we investigate for the first time the performance of an SVM algorithm for predicting TCP after SBRT for stage I NSCLC based on a large multi-institutional database. Our hypothesis was thereby that the more flexible SVM would lead to improvements over a "traditional" multivariate logistic TCP model.

## Methods and Materials

### Patient characteristics

This analysis is based on a cohort of 582 patients with stage I NSCLC who were treated at 13 German and Austrian institutions between 1998 and 2011 as described recently by Guckenberger et al. (11). In the current analysis we used 399 patients with detailed information of tumor stage (clinical stage IA or IB) and a minimum follow-up time of 6 months. Forty-nine (12%) of these patients had a local recurrence after 6 months of follow-up. This was used as ground truth during classification. For the sake of consistency and because most nominal variables were unknown for a large fraction of patients, we decided to restrict analysis to continuous variables. To reduce collinearity among the dosimetric features, analysis was further restricted to considering only biologically effective doses at the isocenter ($BED_{ISO}$) and planning target volume (PTV) periphery ($BED_{PTV}$) which have been shown to be important predictors of TCP (12, 13). This resulted in a total of 7 potential predictors, which are summarized in Table 1. BEDs were calculated based on the LQ formalism as $BED_{ISO/PTV} = n \cdot d_{ISO/PTV}\left(1 + \frac{d_{ISO/PTV}}{\alpha/\beta}\right)$, where n denotes the number of fractions, $d_{ISO/PTV}$ the dose per fraction to the isocenter

**Table 1** Patient characteristics

| Feature | No. of patients | Local control | Local recurrence | $P$ value |
|---|---|---|---|---|
| Age (y): median (range) | 399 | 72 (31-92) | 73 (50-85) | .58 |
| Baseline KI: median (range) | 373 | 80 (40-100) | 80 (60-100) | .80 |
| Baseline FEV1 (l): mean ± SEM | 335 | 1.58 ± 0.04 | 1.81 ± 0.13 | .17 |
| Baseline FEV1%: mean ± SEM | 320 | 60.4 ± 1.4 | 68.1 ± 4.2 | .11 |
| Maximum tumor diameter (cm): median (range) | 210 | 2.5 (0.8-4.8) | 2.9 (1.1-4.7) | .11 |
| $BED_{PTV}$ (Gy): mean ± SEM | 399 | 94.9 ± 1.4 | 83.2 ± 3.4 | .007 |
| $BED_{ISO}$ (Gy): mean ± SEM | 399 | 172.3 ± 2.8 | 141.2 ± 5.8 | .0001 |

*Abbreviations:* $BED_{ISO}$ = biologically effective dose at the isocenter; FEV1 = forced expiratory volume in 1 second; KI = Karnofsky index; SEM = standard error of the mean.
BEDs are calculated with $\alpha/\beta$ = 10 Gy (see text for details).

or PTV periphery, respectively, and $\alpha/\beta$ is assumed to be 10 Gy. The wide variation of fractionation schemas resulted in a broad range of BEDs. For example, the minimum $BED_{ISO}$ of 48 Gy corresponded to 5 × 6 Gy (100%) in 1 patient, and the maximum of 262.5 Gy was achieved with schedules of 3 × 20 Gy (80%) in 15 patients or 3 × 15 Gy (60%) in 7 patients, respectively. The most frequently used fractionation schemas were 3 × 15 Gy (65%) in 26.6% of the patients and 3 × 12.5 Gy (65%) in 20.8% of the patients.

## Statistical tests

Differences between patients with and without local control were assessed through the Wilcoxon rank sum test. Receiver operating characteristic (ROC) curves were compared with the method of DeLong et al (20) using the R package pROC (21). $P$ values <.05 were considered significant. When averaging results, we report the mean and its standard error. For the classification, if a continuous variable was missing for a patient, we assigned the median of its available values.

## The SVM classifier

An SVM is a supervised machine learning classifier that is trained on a set of data with known class labels to subsequently classify data with unknown class membership. During training, the SVM maps the input data into a multidimensional feature space, where it separates 2 classes by finding a linear decision boundary with a maximum margin around it that ideally contains no members of the 2 classes (22, 23). The decision boundary may correspond to a nonlinear boundary in the original data space. In this work, mapping of the input data vectors $x$ is done by a radial basis function kernel with a single parameter $\gamma$ that must be specified before training:

$$K(x_i,x_j)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right) \tag{1}$$

To deal with the problem of regularization for noisy data, a user-specified cost parameter C is introduced that acts to soften the margin. The cost parameter controls the trade-off between allowing transgression of data points across the margin edges toward the other class and a more complex boundary that might lead to overfitting. The representative members of the 2 classes that have transgressed across the margin edges are called the support vectors, and they define the decision boundary.

We used a Java implementation of the freely available software package LIBSVM (24) together with a self-written routine for data reading, preparation, processing, and output. The original version of this classifier was developed for classification of astronomical objects (25). The positive class was defined as local failure. Before training, all continuous data were preprocessed by standardizing to zero mean and unit variance. The same offset and scaling were applied to the test data. Our SVM implementation delivers output probabilities for belonging to a particular class that can be used to order the data and construct ROC curves (26).

## Evaluation of classifier performance

Performance of the classifier was based on area under the ROC curve (AUC). To estimate the test performance of the SVM (ie, the performance expected on an independent test set), we used stratified 10-fold cross-validation (CV). We randomly split the positive and negative examples from our dataset into 10 disjunct parts, each time combining a positive and negative part to yield 1 fold with roughly the same proportion of positive to negative instances as in the original dataset. Each of the 10 folds was then used in turn as a test set, and the remaining 9 folds were used to build the training set as described below. For each of these training-evaluation runs the (C,$\gamma$) pair yielding the largest AUC was identified separately using grid search over the range $C = 2^{-2},2^{-1},\ldots,2^{14}$ and $\gamma = 2^{-13},2^{-12},\ldots,2^{3}$. The best classification results from each training-evaluation run were pooled together to generate a ROC curve according to the procedure given in Fawcett (26).

## Training set modification

Stratified CV creates training sets with a class distribution resembling that of the whole sample, in our case containing about 7 times more negative than positive instances. Inasmuch as such an unbalanced class distribution in the training set of a supervised learning algorithm can lead to poor classification performance (27), we tested 2 strategies to correct for it. The first strategy was undersampling (US), which randomly removes instances from the negative class until their number matches that of the positive examples. US results in fast computation times, but it has the drawback of losing potentially valuable information contained in the removed examples. We therefore investigated the synthetic minority over sampling technique (SMOTE) introduced by Chawla et al (28) as a second strategy. With SMOTE, the minority class is oversampled to a degree $N$ by creating for each instance of the minority class new synthetic instances in the data space along the trajectories to $N$ out of 5 randomly chosen nearest neighbors. We used $N=1$, so for each positive class member in the training set we randomly chose 1 of its 5 nearest positive neighbors, subtracted their input feature vectors, and multiplied each entry in the resulting vector with a uniform random number between 0 and 1 to create a new synthetic positive example in data space. Contrary to simply doubling the negative instances in the training set, SMOTE should lead to a denser distribution of positive instances and a better-defined decision boundary (27). Finally, we randomly undersampled the negative instances until their number equaled the sum of the positive and synthetic examples. Therefore, the combination of SMOTE and US conserves more negative instances than plain US.

## Input feature selection

To identify the variables with the most predictive power, we used a feature forward selection algorithm similar to that described by Chen et al (16). Briefly, parameters were added and eventually replaced successively as input features to train the SVM, each time evaluating the AUC with 10-fold CV as described above, until the best SVM model in terms of AUC had been built.

Because the selection of features could depend on the random number used to split the sample into 10 different folds, we repeated the feature selection procedure 10 times with different random number seeds. To eliminate features that might have been selected as a result of overfitting for a particular division into training/testing sets, the final classifier included only those features that were selected in at least 5 of the 10 feature selection runs. The final classifier was evaluated by randomly changing the patient assignment into training/testing groups 100 times and computing the average AUC, sensitivity, specificity, and accuracy with 10-fold CV as described before. The ROC points from each of the 100 trials were linearly interpolated to 150 points, so that an averaged and smoothed ROC curve could be computed for graphic display.

## Comparison with a standard TCP modeling technique

Other studies have used simple dose cutoffs (10, 11) or a logistic TCP model (5, 29, 30) to classify patients into high-risk and low-risk populations. We compare our SVM-based classification with the latter technique by means of the AUC. A class of multivariate logistic TCP models was defined by

$$\mathrm{TCP}=\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i x_i\right)\div\left[1+\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i x_i\right)\right] \qquad (2)$$

where $x_i$ is one out of $p$ predictors ($1 \leq p \leq 7$) and the $\beta_i$ are regression coefficients. The significance of an individual predictor in a logistic model was assessed through the $t$ statistic based on the residual standard error of its coefficient. The full multivariate model fitted to all patient data was used to check for multicollinearity among the input features using variance inflation factors. We then determined the best set of predictors by forward feature selection based on the Akaike Information Criterion (AIC) using all available patients, and we used 10-fold CV to estimate the test performance of this best model (23).

## Results

Patients who experienced local recurrence received significantly less $\mathrm{BED_{PTV}}$ and $\mathrm{BED_{ISO}}$ (Table 1). Consistently, $\mathrm{BED_{PTV}}$ ($P$=.02) and $\mathrm{BED_{ISO}}$ ($P$=.0001) were significant predictors in univariate logistic regression, whereas only $\mathrm{BED_{ISO}}$ was significantly associated with local control in the full multivariate logistic regression model (Eq. 2) with all 7 features ($P$=.001). All variance inflation factors were below 2.5, indicating no strong collinearity in the data. The best logistic regression model was obtained with $\mathrm{BED_{ISO}}$ and FEV1 as predictors yielding AIC = 281.99.

The cross-validated test performances of this bivariate model and of the 2 univariate logistic models are given in Table 2. The AUC of the former, to which we will simply refer as "the logistic TCP model" in the following, was 0.680 ± 0.040 with an optimal TCP cutoff at 87.1%, yielding sensitivity and specificity of 63.3% and 66.0%, respectively (64.7% accuracy). It is displayed graphically in Figure 1, with parameters being the averaged ("cross-validated") maximum likelihood parameters from each of the 10 CV training runs. This model suggests increasing TCP with increasing $\mathrm{BED_{ISO}}$ and, in particular at low $\mathrm{BED_{ISO}}$, decreasing FEV1.

The results of the SVM feature selection are presented in Table 3 for both strategies of creating balanced training sets, namely, US and SMOTE + US. The corresponding ROC curves are shown in Figure 2 together with the ROC curve obtained from the best logistic TCP model. The mean AUC, sensitivity, specificity, and accuracy taken over all 10 feature selection trials are also given in Table 3. For all SVM classifiers, the AUC was larger than that of the logistic model, but the differences were statistically significant only when the combination of SMOTE and US was used ($P$=.012 ± .003). The better classification performance coincided with more selected input features, probably reflecting a wider variability in the feature space caused by more training data (see also supplementary online material, available at www.redjournal.org). The superior performance of SMOTE + US compared with plain US motivated us to use this same strategy for the final classifier, which was built with the following features that were selected by at least 5 of the 10 SVMs used for feature selection (Table 3): $\mathrm{BED_{ISO}}$, age, baseline Karnofsky index (KI), and FEV1. The selection of these features was insensible against using 2-fold CV instead of 10-fold CV.

Table 2 contains the results for the final classifier with SMOTE + US. The small standard error of the mean AUC indicates that the performance of the final SVM was insensitive to the particular data splitting into the CV folds. In Figure 3 we have plotted the smoothed and averaged ROC curve of the final SVM model together with the ROC curve of the logistic TCP model. Comparison of both curves showed that the SVM model's AUC was significantly larger than that of the TCP model ($P$= .032 ± .002). By use of an output probability threshold at 50%, the average sensitivity and specificity for the SVM were 67.0% and 78.7%, respectively (72.8% accuracy).

## Discussion

In this work we have shown that an SVM significantly improves TCP predictions compared with a logistic dose−effect model by taking into account additional treatment plan and patient characteristics besides BED. Together with the fruitful attempts of SVM-based prediction of severe radiation-induced pneumonitis undertaken by Chen et al (16), Das et al (31), and Naqa et al (14), our investigation suggests that SVMs and other machine-learning methods could be valuable tools leading to improved prediction of

**Table 2** Performance of 3 logistic TCP models and the final SVM classifier

| Classifier | Parameters | AUC | TCP cutoff (%) | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|---|---|---|---|---|---|---|
| Logistic model | $BED_{PTV}$ | 0.610 ± 0.043 | 87.3 | 77.6 | 46.6 | 62.1 |
| Logistic model | $BED_{ISO}$ | 0.662 ± 0.037 | 87.1 | 63.3 | 64.9 | 64.1 |
| Logistic model | $BED_{ISO}$, FEV1 | 0.680 ± 0.040 | 87.1 | 63.3 | 66.0 | 64.7 |
| SVM | $BED_{ISO}$, FEV1, KI, age | 0.789 ± 0.001 | 50 | 67.0 ± 0.5 | 78.7 ± 0.3 | 72.8 ± 0.2 |

*Abbreviations:* AUC = area under the curve; $BED_{ISO}$ = biologically effective dose at the isocenter; FEV1 = forced expiratory volume in 1 second; KI = Karnofsky index; SMOTE = synthetic minority over sampling technique; SVM = support vector machine; TCP = tumor control probability; US = undersampling.

All performance measures are based on 10-fold cross-validation and are given as mean ± standard error of the mean (SEM). SEMs are calculated from 100 trials with different splits of training and test sets in case of the SVM or (for AUC only) by the method of DeLong et al (20) in case of the logistic TCP models.

high-risk and low-risk groups, thus helping the clinician to balance the benefit of the treatment with the anticipated risk.

Traditional radiobiological modeling based on mathematical descriptions of a TCP function has some shortcomings that are circumvented by using machine learning approaches. First and foremost, interactions between influencing parameters are usually complicated and are not known a priori, leaving the investigator with a trial-and-error approach in the attempt to incorporate them into formulas that model dose−effect relationships. In this regard, parametric models are often limited by their small number of parameters and by assumptions about correlations among variables such as linearity, as was the case in the logistic TCP model we investigated. Second, parametric models are not always straightforward to interpret. Good examples are logistic TCP models, as we can easily illustrate with our data: With an univariate model containing only $BED_{ISO}$ as a predictor, we obtain a cross-validated maximum likelihood value of 1.4 Gy for a regression parameter usually called TCD50 and interpreted as "the dose to achieve a 50% TCP" (29). Clearly, this interpretation is not valid in our case because our data do not span that range (see also discussion in Ref.5). This example highlights the advantages of parameter-free classification approaches such as the SVM, which solely uses the data at hand to find an optimal nonlinear separation between 2 classes in the space spanned by the variables of interest and whose results are easily interpretable in terms of outcome probabilities. Finally, machine learning methods usually perform better than traditional logistic regression on small datasets (14).

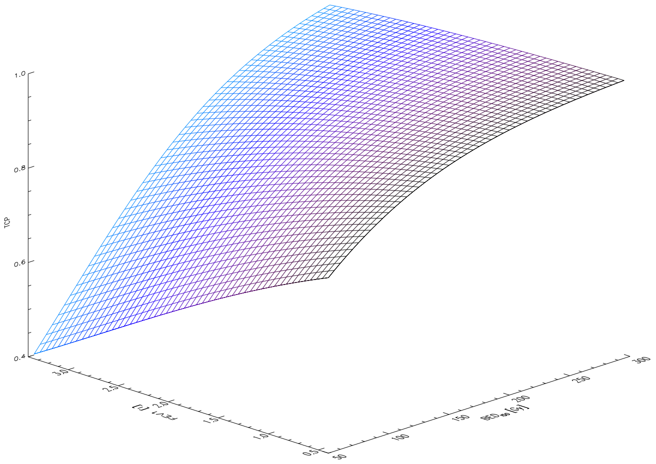

**Fig. 1.** Maximum likelihood fit of the logistic tumor control probability model. The cross-validated maximum likelihood parameters were $(\beta_0,\ \beta_{BED_{ISO}},\ \beta_{FEV1}) = (-0.764,\ -0.0134\,\mathrm{Gy}^{-1},\ 0.5381^{-1})$.

Our selection of BED as an input feature is an example of applying a priori knowledge about interaction between variables—in this case, number of fractions and fraction dose—as part of the data selection step with the goal of increasing data heterogeneity and improving predictive power (9). Previous studies have clearly established the importance of both $BED_{ISO}$ and $BED_{PTV}$ for determining TCP (12), whereas our results suggest that $BED_{ISO}$ might be the better surrogate for the dose actually delivered to the tumor in this study cohort, consistent with the findings of Guckenberger et al (11). We further showed that taking into account age, baseline KI, and FEV1 improves predictive performance, although none of these variables would be useful as a sole predictor.

Baseline KI was always selected next to $BED_{ISO}$ by the SVM but played no role in multivariate logistic modeling, which points toward a higher-order interaction between these 2 variables. By contrast, our data implicated a decrease in the pulmonary function indicator FEV1 as a predictor of higher TCP at a given $BED_{ISO}$ (Fig. 1). For example, assuming all patients receiving $BED_{ISO}$ = 150 Gy, TCP would be predicted as 93% for those with

**Table 3** Results of the feature selection

| | US | SMOTE + US |
|---|---|---|
| Features | $BED_{ISO}$ (10)<br>Baseline KI (8)<br>FEV1 (4)<br>Age (3)<br>Maximum tumor diameter (2) | $BED_{ISO}$(10)<br>Baseline KI (10)<br>FEV1 (7)<br>Age (6)<br>$BED_{PTV}$ (2)<br>FEV1% (2)<br>Maximum tumor diameter (1) |
| AUC | 0.725 ± 0.005 | 0.811 ± 0.004 |
| Sensitivity | 69.0 ± 1.5 | 69.6 ± 2.3 |
| Specificity | 65.3 ± 1.3 | 80.1 ± 1.6 |
| Accuracy | 67.2 ± 0.8 | 74.9 ± 0.7 |

*Abbreviations:* AUC = area under the curve; $BED_{ISO}$ = biologically effective dose at the isocenter; FEV1 = forced expiratory volume in 1 second; KI = Karnofsky index; $BED_{PTV}$ = biologically effective dose at the planning target volume periphery; SMOTE = synthetic minority over-sampling technique; US = undersampling. The number in brackets gives the number of times the specific feature was selected in 10 feature selection trials (see text for details).

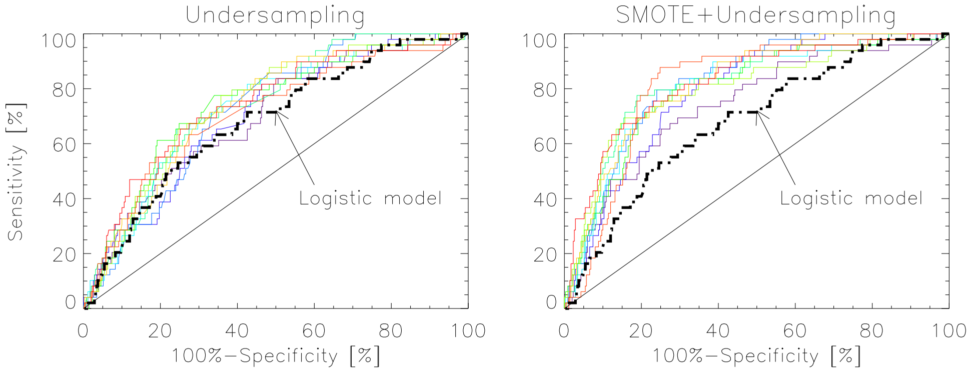

**Fig. 2.** Receiver operating characteristic (ROC) curves for the classifiers that were evaluated for input feature selection. Each panel shows the ROC curves that resulted from 10 different trials that we ran as part of the feature selection algorithm (see text for details). Left, the training set was balanced using plain undersampling. Right, we used a combination of the synthetic minority over sampling technique (SMOTE) and undersampling. Dash-dotted black line = ROC curve for the logistic tumor control probability model. Solid diagonal = performance of a random classifier with area under the curve = 0.5.

a FEV1 of 0.4 l but only 78% for those with 2.81. It is beyond the scope of this paper to investigate whether this represents some real physiologic effect. Pulmonary function could principally influence tumor control through its effect on breathing motion during irradiation. Stevens et al (32) found no correlation between FEV1 and tumor motion, but their study was limited by the reliance on orthogonal 2-dimensional radiographic plates and a small and heterogeneous patient cohort. Keeping in mind that FEV1 was unknown for 16% of our patients, we cautiously acknowledge that FEV1 seems to influence TCP in this large patient cohort, but further studies are needed to confirm these findings.

Although tumor size is an established prognostic factor for local recurrence in the literature (3, 4), the maximum tumor diameter was chosen only 4 out of 10 times as an input feature and therefore was not used in the final SVM. Patients with local recurrence had larger tumors, but this difference was not statistically significant (Table 1). The fact that tumor size was known for only 210 of the 399 patients, whereas the other patients were assigned the median tumor diameter, likely downgraded the importance of this feature.

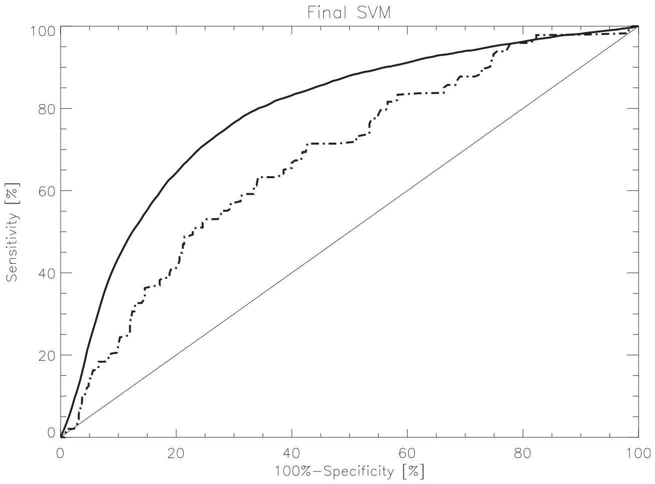

**Fig. 3.** Receiver operating characteristic curves of the final support vector machine (SVM) classifier (solid line) and the logistic tumor control probability (TCP) model (dash-dotted line). The area under the curve (AUC) of the SVM classifier was significantly larger than that of the logistic TCP model, which modeled TCP as a function of biologically effective dose at the isocenter and forced expiratory volume in 1 second ($P$<.05). Solid straight line = a random classifier having AUC = 0.5.

Considering the clinical relevance of our findings, 2 issues need to be discussed. First, CV is an established way of estimating classifier performance on independent test data, but ultimately such data would be needed to further validate our SVM. Our multi-institutional database could be seen as an advantage in this regard, because it already includes interinstitutional variance in the training data to some extent. However, it also seems clear that more cases of tumor recurrence after SBRT should be collected to deal with the class imbalance problem that we addressed by creating synthetic patients of the positive class through SMOTE and by undersampling the negative class. Second, given the large number of potential TCP-influencing variables, predictions from classifiers that were trained only on a subset of such variables should be interpreted not as good estimates of a patient's true TCP but more as a way to group patients into high-risk and low-risk groups. Unfortunately, dose−volume metrics were not available, and we decided to exclude other putative factors influencing treatment outcome that were unknown for many patients, such as tumor histology and location (central vs peripheral). It would be important to investigate these and other factors in future analyses to improve predictions further. Specifically, markers of tumor cell and host metabolism such as positron emission tomography scans using $^{18}$F-fluoro-2-deoxy-glucose (33) or blood glucose levels (34) have emerged as potent predictors of tumor control and survival in a variety of cancers. Finally, Perez et al (35) have recently demonstrated that the response of NSCLC to fractionated radiation therapy depends on p53 status, raising the question whether genotyping tumors will help to improve treatment outcome predictions in the future (36).

## References

1. Onishi H, Araki T, Shirato H, et al. Stereotactic hypofractionated high-dose irradiation for stage I nonsmall cell lung carcinoma: Clinical outcomes in 245 subjects in a Japanese multiinstitutional study. *Cancer* 2004;101:1623-1631.
2. Wulf J, Baier K, Mueller G, et al. Dose-response in stereotactic irradiation of lung tumors. *Radiother Oncol* 2005;77:83-87.

3. Onimaru R, Fujino M, Yamazaki K, et al. Steep dose-response relationship for stage I non-small-cell lung cancer using hypofractionated high-dose irradiation by real-time tumor-tracking radiotherapy. *Int J Radiat Oncol Biol Phys* 2008;70:374-381.
4. Guckenberger M, Wulf J, Mueller G, et al. Dose-response relationship for image-guided stereotactic body radiotherapy of pulmonary tumors: Relevance of 4D dose calculation. *Int J Radiat Oncol Biol Phys* 2009;74:47-54.
5. Ohri N, Werner-Wasik M, Grills IS, et al. Modeling local control after hypofractionated stereotactic body radiation therapy for stage I non-small cell lung cancer: A report from the elekta collaborative lung research group. *Int J Radiat Oncol Biol Phys* 2012;84:e379-e384.
6. Choi NC, Fischman AJ, Niemierko A, et al. Dose-response relationship between probability of pathologic tumor control and glucose metabolic rate measured with FDG PET after preoperative chemoradiotherapy in locally advanced non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys* 2002;54:1024-1035.
7. Ruggieri R, Stavreva N, Naccarato S, et al. Computed 88% TCP dose for SBRT of NSCLC from tumour hypoxia modelling. *Phys Med Biol* 2013;58:4611-4620.
8. El Naqa I. Machine learning methods for predicting tumor response in lung cancer. *WIREs Data Mining Knowl Discov* 2012;2:173-181.
9. Lambin P, Van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology: Multifactorial decision support systems. *Nat Rev Clin Oncol* 2013;10:27-40.
10. Hof H, Muenter M, Oetzel D, et al. Stereotactic single-dose radiotherapy (radiosurgery) of early stage nonsmall-cell lung cancer (NSCLC). *Cancer* 2007;110:148-155.
11. Guckenberger M, Allgäuer M, Appold S, et al. Safety and efficacy of stereotactic body radiotherapy for stage I non-small-cell lung cancer in routine clinical practice: A patterns-of-care and outcome analysis. *J Thorac Oncol* 2013;8:1050-1058.
12. Chi A, Tome WA, Fowler J, et al. Stereotactic body radiation therapy in non−small-cell lung cancer. *Am J Clin Oncol* 2011;34:432-441.
13. Senthi S, Haasbeek CJA, Slotman BJ, et al. Outcomes of stereotactic ablative radiotherapy for central lung tumours: A systematic review. *Radiother Oncol* 2013;106:276-282.
14. Naqa IE, Deasy JO, Huang E, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol* 2010;49:1363-1373.
15. Kotsiantis SB. Supervised machine learning: A review of classification techniques. *Informatica* 2007;31:249-268.
16. Chen S, Zhou S, Yin F, et al. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *Med Phys* 2007;34:3808-3814.
17. Oh JH, Craft J, Al Lozi R, et al. A Bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol* 2011;56:1635-1651.
18. Niméus-Malmström E, Krogh M, Malmström P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence after breast-conservation surgery, with or without postoperative radiotherapy. *Breast Cancer Res* 2008;10:R34.
19. Wan X, Zhao Y, Fan X, et al. Molecular prognostic prediction for locally advanced nasopharyngeal carcinoma by support vector machine integrated approach. *PLoS ONE* 2012;7:e31989.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988;44:837-845.
21. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
22. Cortes C, Vapnik VN. Support vector networks. *Machine Learning* 1995;20:273-297.
23. James G, Witten D, Hastie T, et al. An introduction to statistical learning with applications in R. New York: Springer Science+Business Media; 2013.
24. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Sys Techn* 2011;2. Software available at, http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. Last access was December, 2013.
25. Klement RJ, Bailer-Jones CAL, Fuchs B, et al. Classification of field dwarfs and giants in rave and its use in stellar stream detection. *Astrophys J* 2011;726:103.
26. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters* 2006;27:861-874.
27. Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: Boulicaut J-F, Esposito F, Giannotti F, et al., editors. Machine Learning: ECML 2004Vol 3201. Berlin, Heidelberg: Springer-Verlag; 2004. p. 39-50.
28. Chawla NV, Bowyer RW, Hall LO, et al. SMOTE: Synthetic minority over-sampling technique. *J Art Intell Res* 2002;16:321-357.
29. Okunieff P, Morgan D, Niemierko A, et al. Radiation dose-response of human tumors. *Int J Radiat Oncol Biol Phys* 1995;32:1227-1237.
30. Stuschke M, Pottgen C. Altered fractionation schemes in radiotherapy. *Frontiers Oncol* 2010;42:150-156.
31. Das SK, Chen S, Deasy JO, et al. Combining multiple models to generate consensus: Application to radiation-induced pneunomitis prediction. *Med Phys* 2008;35:5098-5109.
32. Stevens CW, Munden RF, Forster KM, et al. Respiratory-driven lung tumor motion is independent of tumor size, tumor loction, and pulmonary function. *Int J Radiat Oncol Biol Phys* 2001;51:62-68.
33. Miles KA, Williams RE. Warburg revisited: Imaging tumour blood flow and metabolism. *Cancer Imaging* 2008;8:81-86.
34. Klement RJ, Kämmerer U. Is there a role for carbohydrate restriction in the treatment and prevention of cancer? *Nutr Metab* 2011;8:75.
35. Perez BA, Ghafoori AP, Lee C, et al. Assessing the radiation response of lung cancer with different gene mutations using genetically engineered mice. *Frontiers Oncol* 2013;3:72.
36. Hirst DG. Molecular biology: The key to personalised treatment in radiation oncology? *Brit J Radiol* 2010;83:723-728.