World J Surg (2012) 36:830–835
DOI 10.1007/s00268-012-1449-8

World Journal of Surgery

# A Mathematical Model for Predicting Malignancy of Solitary Pulmonary Nodules

Yun Li · Jun Wang

Published online: 2 February 2012

## Abstract

*Background* The goal of the present study was to differentiate between benign and malignant solitary pulmonary nodules (SPN) by developing a mathematical prediction model.

*Methods* Records from 371 patients (197 male, 174 female) with SPN between January 2000 and September 2009 were reviewed (group A). Clinical data were collected to estimate the independent predictors of malignancy of SPN with multivariate logistic regression analysis. A clinical prediction model was subsequently developed. Between October 2009 and May 2011, data from an additional 145 patients with SPN were used to validate this new clinical prediction model (group B). The same data were also estimated with two previously published models for comparison with our new model.

*Results* The median patient age was 57.1 years in group A; 54% of the nodules were malignant and 46% were benign. Logistic regression analysis identified six clinical characteristics (age, diameter, border, calcification, spiculation, and family history of tumor) as independent predictors of malignancy in patients with SPN. The area under the receiver operator characteristic (ROC) curve for our model ($0.874 \pm 0.028$) was higher than those generated using the other two reported models. In our model, sensitivity = 94.5%, specificity = 70.0%, positive predictive value = 87.8%, and negative predictive value = 84.8%).

*Conclusions* Age, diameter, border, calcification, spiculation, and family history of tumor were independent predictors of malignancy in patients with SPN. Our prediction model was sufficient to estimate malignancy in patients with SPN and proved to be more accurate than the two existing models.

## Introduction

Solitary pulmonary nodules (SPN), usually incidentally discovered on a chest radiograph or chest computed tomography (CT) scan performed for any reason are the most common manifestation of lung cancer [1]. Detecting and diagnosing SPN is critical, as early identification of malignant nodules is crucial to the chance for successful treatment. Solitary pulmonary nodules, however, are usually small, located deep in the pulmonary parenchyma, and often yield atypical imaging findings. While researchers are seeking advanced image techniques [2–6], more and more clinicians have stressed that in addition to the radiographic and image characteristics, such as size, edge, or contour, clinical information such as age and smoking history should also be factored into the decision of whether a nodule is benign or malignant. Confronting such massive data, clinical experience and judgment may not be reproducible or reliable, whereas a quantitative model might have advantages in accuracy and reproducibility, will not be uninfluenced by personal judgment, and can provide outcome exchange ability.

With these advantages in mind, researchers began to suggest that a clinical prediction equation has the potential to facilitate clinical decision making [7–12]. One widely cited prediction model was proposed by Swensen et al. [8, 9] at the Mayo Clinic, who have reported that the capability of a mathematical model in judging the characteristics of SPN is similar to the clinical judgment of experienced physicians. Nevertheless, one of the

Y. Li (✉) · J. Wang
Department of Thoracic Surgery, People's Hospital of Peking University, No. 11 Xizhimen South Street, Xicheng District, 100044 Beijing, People's Republic of China
e-mail: xiongwai@263.net

limitations in their model is that 12% of the patients did not have a final diagnosis. Recently, Gould et al. [10] at Department of Veterans Affairs (VA) Cooperative Study reported an excellent agreement between the predicted probability and the observed frequency of malignant SPN. Their area under the curve of the receiver operator characteristic (ROC) is 0.79 (95% Confidence Interval [95% CI], 0.74–0.84). However, their model requires external validation in an independent cohort of patients with SPN, their clinical predictors are relatively small, and they were unable to evaluate spiculation and a remote history of extrathoracic cancer.

To improve the accuracy of the model, systematic and comprehensive clinical and imaging data collection with specific diagnoses is needed. Mery et al. [13] reported that age, tumor size, smoking history, and time of smoking cessation were independent factors, and Zhang [14] reported patient age, tumor size, and CT imaging were independent factors. The differences in the independent factors emphasize the significance of the source of the data (i.e., foreign versus domestic clinical data). In the present study, we aimed to develop a prediction model for patients with SPN based on a comprehensive data collection and thorough analysis.

## Materials and methods

### Clinical data

Between January 2000 and September 2009, medical records from 405 patients with a radiographic diagnosis of SPN were reviewed. Of these, 9 were excluded because data were incomplete and 25 were excluded because of a history of pulmonary or extrapulmonary malignancy in 5 years. In total, 371 cases were enrolled to create a mathematical model (group A). Clinical data collected include age of the patient, gender, smoking history, history of cancer, family history of cancer, calcification, spiculation, lobulation, pleural retraction, clear border, cavity, vascular convergence, tumor site (upper lobe or lower lobe, left or right), and diameter of tumor.

Clinical data were also collected from an additional 150 patients with a radiographic diagnosis of SPN between October 2009 and May 2011. Of these, 5 cases were excluded because of a history of pulmonary or extrapulmonary malignancy in 5 years. In total, 145 cases were enrolled to test the constructed mathematical model (group B).

### Surgical procedure

All patients included in groups A and B underwent surgical resection of pulmonary nodules, after which a definitive pathologic diagnosis of an SPN as benign or malignant was established. Surgical procedures included tumor enucleation, pulmonary wedge resection, and lobectomy.

### Statistical analysis

SPSS13.0 software (2004, IBM, Armonk, NY) was used for statistical analysis. For single factor analysis, the information for group A was analyzed to identify all factors affecting the probability of malignancy for SPN. Multivariate logistic regression was then performed to select independent predictive factors. A mathematical model for SPN was subsequently devised based on the results of the multivariate logistic regression. Receiver operator characteristic curves were created and the areas under the curves were calculated. Appropriate cut-off points were determined and the sensitivity, specificity, positive predictive value, and negative predictive value were calculated. A $p$ value $< 0.05$ was considered statistically significant.

## Results

Pathology revealed 142 cases of benign disease (38.3%) and 229 cases of malignant disease (61.7%) in group A, and 47 cases of benign disease (32.4%) and 98 cases of malignant disease (67.6%) in group B. There was no significant statistical difference in age, sex, or nodule diameter between the two groups ($p > 0.05$).

### Univariate and multivariate analyses

The univariate analysis results are shown in Table 1. There were significant differences in age, smoking history and smoking, family history of cancer, calcification, spiculation, lobulation, pleural retraction sign, clear border, and maximum tumor diameter between the benign and malignant SPN patients. Of these, age, family history of cancer, spiculation, calcification, clear border of SPN, and maximum tumor diameter in benign and malignant SPN were identified as independent risk factors for malignant and benign SPN through multivariate logistic regression analysis (Table 2).

### Model construction

The following formula was employed to describe the malignant probability:

$$\begin{aligned} p = e^x/(1+e^x), x = &-4.496 + (0.07 \times \text{Age}) \\ &+ (0.676 \times \text{diameter}) + (0.736 \times \text{spiculation}) \\ &+ (1.267 \times \text{family history of cancer}) \\ &- (1.615 \times \text{calcification}) - (1.408 \times \text{border}), \end{aligned}$$

**Table 1** Univariate analysis of data collected from patients included in group A

| | Benign | Malignant | *p* Value |
|---|---|---|---|
| Age, years | 48.0 ± 14.6 | 61.2 ± 13.1 | <0.001 |
| Diameter, cm | 1.72 ± 0.76 | 2.13 ± 0.73 | <0.001 |
| Gender | | | |
| Male | 72 (50.7%) | 125 (54.6%) | 0.467 |
| Female | 70 (49.3%) | 104 (45.4%) | |
| Smoking history | | | |
| No | 96 (67.6%) | 119 (52.0%) | 0.003 |
| Smoking quantity, pieces/year | 169.8 ± 328.3 | 260.6 ± 410.3 | 0.026 |
| Family history of cancer | | | |
| No | 135 (95.1%) | 200 (87.3%) | 0.014 |
| Previous cancer history | | | |
| No | 140 (98.6%) | 222 (96.9%) | 0.512 |
| Calcification | | | |
| No | 124 (87.3%) | 220 (96.1%) | 0.002 |
| Spiculation | | | |
| No | 124 (87.3%) | 149 (65.1%) | <0.001 |
| Lobulation | | | |
| No | 123 (86.6%) | 170 (74.2%) | 0.004 |
| Pleural retraction sign | | | |
| No | 123 (86.6%) | 165 (72.1%) | 0.001 |
| Clear border | | | |
| No | 74 (52.1%) | 197 (86.0%) | <0.001 |
| Cavity | | | |
| No | 128 (90.1%) | 202 (88.2%) | 0.564 |
| Vascular convergence | | | |
| No | 136 (95.8%) | 209 (91.3%) | 0.098 |
| Position | | | |
| Upper lobe | 67 (47.2%) | 131 (57.2%) | 0.06 |
| Lower lobe | 75 (52.8%) | 98 (42.8%) | |
| Left side | 63 (44.4%) | 94 (41.0%) | 0.53 |
| Right side | 79 (55.6%) | 135 (59.0%) | |

where *e* is the natural logarithm, and the value for the last four elements, i.e., family cancer history, calcification, spiculation, and border, equals 1 if the element exists, and 0 otherwise.

The clinical data of the 371 cases included in group A were used in the mathematical model to calculate a predicted probability of malignancy for all patients. A *p* value of 0.463 was ultimately selected as a cut-off point and *p* values > 0.463 should be considered malignant disease and $p < 0.463$ should be considered benign.

## Model validation

Clinical data of the patients in group B were then used to test the accuracy of the model by comparing the calculated result (*p* value) with the pathology results; ROC curves were then created (Fig. 1). The area under the ROC curve was 0.874 ± 0.028. The sensitivity of this model for group B was 94.5%, the specificity was 70.0%, the positive predictive value was 87.8%, and the negative predictive value was 84.8%.

**Table 2** Multivariate logistic regression analysis

| Factor | Regression coefficient | *p* Value | Odds Ratio value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| Age | 0.070 | <0.001 | 1.073 | 1.050 | 1.096 |
| Diameter | 0.676 | <0.001 | 1.966 | 1.379 | 2.803 |
| Family history of cancer | 1.267 | 0.016 | 3.550 | 1.264 | 9.968 |
| Calcification | −1.615 | 0.00 | 0.199 | 0.067 | 0.587 |
| Spiculation | 0.736 | 0.035 | 2.088 | 1.055 | 4.135 |
| Clear border | −1.408 | <0.001 | 0.245 | 0.133 | 0.451 |
| Constant | −4.496 | <0.001 | 0.011 | | |

In addition, two established foreign mathematical models were tested using group B data, the Mayo model [4] and the VA model [5].

### *The mayo model*

Independent factors in the Mayo model were age, smoking history, cancer history, diameter, spiculation, and site in left side. The calculation was based on the formula

$$p = e^x/(1 + e^x)$$

where $x = -6.8272 + (0.0391 \times \text{age}) + (0.7917 \times \text{smoking history}) + (1.3388 \times \text{cancer history}) + (0.1274 \times \text{diameter}) + (1.0407 \times \text{spiculation}) + (0.7838 \times \text{the upper lobe})$. The area under the ROC curve of group B was 0.784 ± 0.038 (Fig. 2).

### *The VA group model*

For the VA model, independent factors were age, smoking, tumor diameter, and time of smoking cessation. The equation was

$$p = e^x/(1 + e^x)$$

where $x = -8.404 + (2.061 \times \text{smoking history}) + (0.779 \times \text{age}) + (0.112 \times \text{diameter}) - (0.567 \times \text{time of smoking cessation})$. The area under the ROC curve of group B was 0.754 ± 0.040 (Fig. 3).

The area under the ROC curve of the mathematical model created in the present study was significantly higher

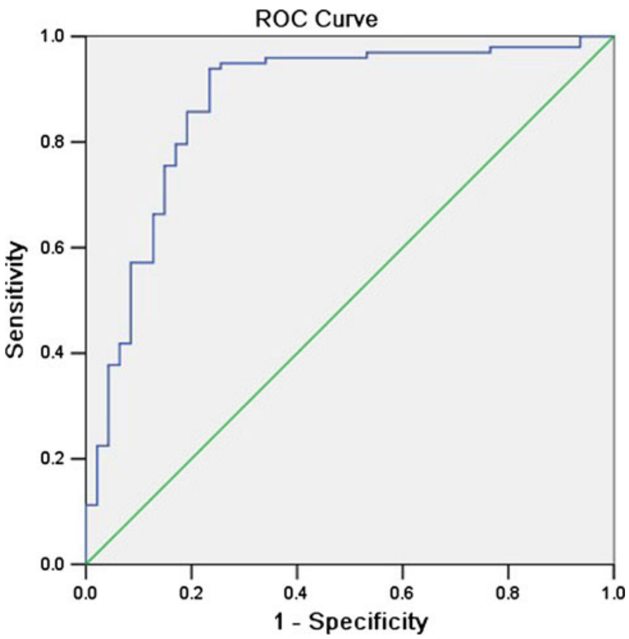

**Fig. 1** Receiver operator characteristic (ROC) curve generated using our proposed model

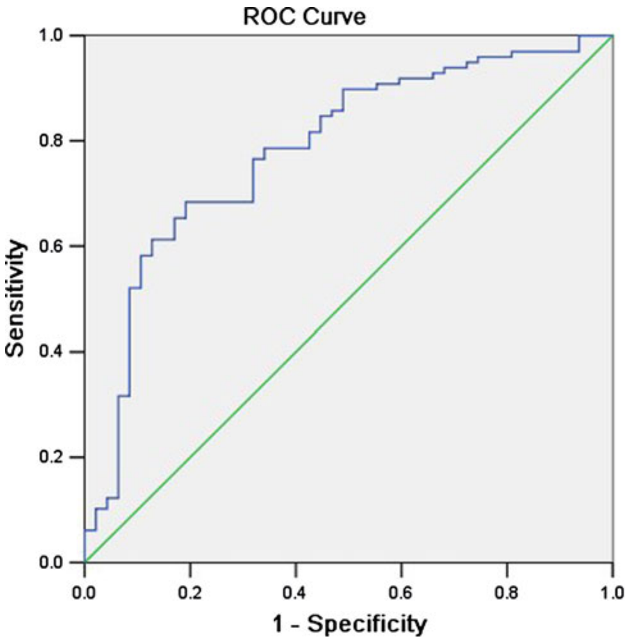

**Fig. 2** Receiver operator characteristic curve generated using the Mayo model

than the other two foreign forecasting models ($p < 0.05$), as described in Table 3 and Figure 4.

## Discussion

Solitary pulmonary nodule evaluation is a mathematical tool of major interest to clinicians. It provides a conventional method for establishing a prediction model after univariate and multivariate regression analysis. Up until now, such mathematical models have all been studied retrospectively. The present work is also a retrospective study; however, to surpass the previous works, it gives the most comprehensive data collection of both clinical and imaging information, with a definitive pathology diagnosis for each patient. This has never been done in any of the studies reported to date, and it cannot be achieved by prospective study. Because all the patients were pathologically diagnosed through surgical treatment, only the retrospective method can get such accurate results. Multivariate regression analysis for modeling is widely used for this kind of research, and we followed this classic method without much improvement. Thus our model is more accurate because it is based on comprehensive and systematic data collection, especially with a clinical pathologic diagnosis. Evidenced by the ROC curves and compared with the widely accepted models, the more comprehensive the data collected, the more reliable the model will be.

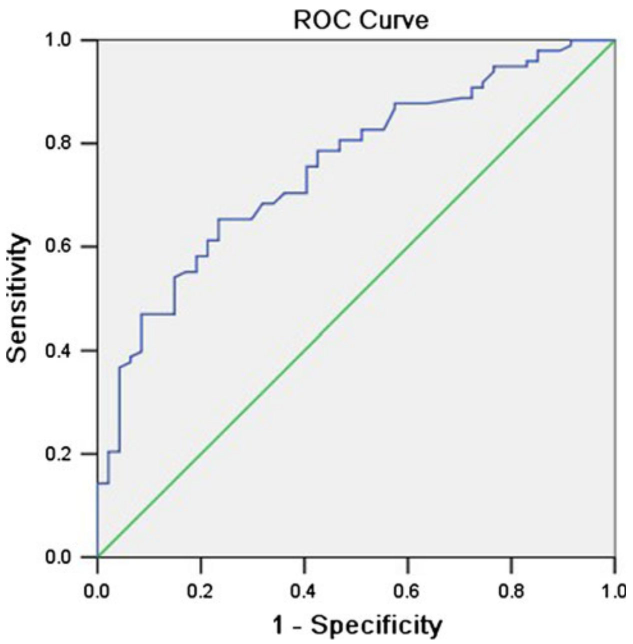

**Fig. 3** Receiver operator characteristic curve generated using the VA model. Diagonal segments are produced by ties

Among the independent factors determined by multivariate analysis, age [8], nodule diameter [12–15], calcification [3–6], spiculation [7], and border [16] have been reported before, but family history of cancer is a new independent factor that has not been studied yet. Although malignant tumor has not been demonstrated to show any specific genetic characteristic, several researchers [17–21] whose studies have been published in the past decade have indicated that tumorigenesis might be a compound result of

**Table 3** Comparison of the receiver operator characteristic (ROC) curves of three mathematical models analyzed in this study

| Model | Area under the curve (AUC) | Standard error | *p* Value | 95% CI | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| This study | 0.874 | 0.028 | 0.000 | 0.809 | 0.923 |
| Mayo model [9] | 0.784 | 0.038 | 0.004 | 0.708 | 0.848 |
| VA model [10] | 0.754 | 0.040 | 0.002 | 0.675 | 0.821 |

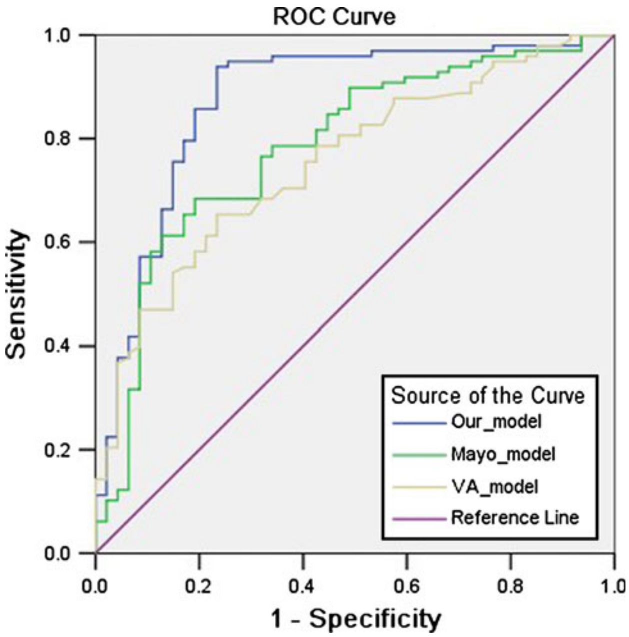

**Fig. 4** Comparison of ROC curves generated using the three models analyzed in this study, the area under ROC curve of our new model is significantly higher than that of the Mayo and VA models. Diagonal segments are produced by ties

genetic and environmental factors; i.e., cancer arises from a genetic predisposition and is stimulated by environment influences. This topic has attracted considerable attention in recent years. We believe that a genetic susceptibility to cancer or a hereditary predisposition to cancer could significantly increase the malignancy probability in patients. It provides us another consideration to evaluate SPN.

In addition, it has been reported [3] that smoking and nodule location are independent factors in malignancy of SPN. In the present study, however, both have been proved to be dependent factors. It is possible that there were more patients with adenocarcinoma in the present study. The relationship between smoking and adenocarcinoma is not as clear as that with squamous cell carcinoma. Alternatively, all of the patients in this study underwent a surgical procedure to obtain a pathologic diagnosis. Some non-smoking SPN patients whose SPN were determined to be benign did not undergo operation and were not included in this study. The ratio of non-smoking patients with benign nodules in our group decreased correspondingly, so whether a history of smoking is significantly associated with malignancy in patients with SPN requires further study. As for nodule location, there were a high number of patients in this group with tuberculosis, which is known to occur in the upper lobe. Also, in China, the incidence of tuberculosis is higher than that in Western countries. This study suggests that there is no significant correlation between nodule location and the probability of malignancy of SPN.

The model was well fitted to the independent cohort of 145 patients seen from October 2009 to May 2011, which is a good demonstration of the accuracy of the prediction model. In addition, we compared our model with two previously published models [9, 10]. For the three models tested, the area under the ROC curve was 0.874, 0.784, and 0.754, and there were significant differences among the three groups. The larger the area under the curve, the more accurate the mathematical model appeared. Based on this assertion, the new model described herein was significantly better than the existing formulae. Moreover, the cut-off value for the model was obtained by calculating the optimum sensitivity and specificity. $p$ values $> 0.463$ should be considered malignant disease and $p < 0.463$ should be considered benign. The sensitivity of this model was 94.5%, the specificity was 70.0%, the positive predictive value was 87.8%, and negative predictive value was 84.8%.

In conclusion, age of patient, diameter, border, calcification, spiculation, and family history of tumor were independent predictors of malignancy in patients with SPN. The devised prediction model was more accurate than two previously described models and was able to predict malignancy in patients with SPN. Although the mathematical models provide an objective basis for judging the character of SPN, it remains a clinical tool in that it cannot be used as a substitute for a pathologic diagnosis. Previous reports have indicated that the possibility of malignancy in patients with SPN is very high (80% or more) [22–24]. As patient age increases, the possibility of malignancy also increases significantly. Therefore, clinicians need to seriously consider all SPN, especially in patients with the six risk factors noted above.

## References

1. Soubani AO (2008) The evaluation and management of the solitary pulmonary nodule. Postgrad Med J 84:459–466
2. Ost D, Fein AM, Feinsilver SH (2003) Clinical practice. The solitary pulmonary nodule. N Engl J Med 348:2535–2542
3. Erasmus JJ, Connolly JE, McAdams HP et al (2000) Solitary pulmonary nodules: part I. Morphologic evaluation for differentiation of benign and malignant lesions. Radiographics 20:43–58

4. Bateson EM (1965) An analysis of 155 solitary lung lesions illustrating the differential diagnosis of mixed tumours of the lung. Clin Radiol 16:51–65
5. Zerhouni EA, Stitik FP, Siegelman SS et al (1986) CT of the pulmonary nodule: a cooperative study. Radiology 160:319–327
6. Mahoney MC, Shipley RT, Corcoran HL et al (1990) CT demonstration of calcification in carcinoma of the lung. AJR Am J Roentgenol 154:255–258
7. Khan A, Herman PG, Vorwerk P et al (1991) Solitary pulmonary nodules: comparison of classification with standard, thin-section, and reference phantom CT. Radiology 179:477–481
8. Swensen SJ, Silverstein MD, Edell ES et al (1999) Solitary pulmonary nodules: clinical prediction model versus physicians. Mayo Clin Proc 74:319–329
9. Swensen SJ, Silverstein MD, Llstrup DM et al (1997) The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med 157:849–855
10. Gould MK, Ananth L, Barnett PG et al (2007) A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest 131:383–388
11. Nakamura K, Yoshida H, Engelmann R et al (2000) Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. Radiology 214:823–830
12. Gurney JW (1993) Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. Radiology 186:405–413
13. Mery CM, Pappas AN, Bueno R et al (2004) Relationship between a history of antecedent cancer and the probability of malignancy for a solitary pulmonary nodule. Chest 125:2175–2181
14. Zhang Y, Ding J-A (2006) Benign and malignant solitary pulmonary nodule: analysis of risk factors. Chin J Clin Thorac Cardiovasc Surg 13:162–165
15. MacMahon H, Austin JH, Gamsu G et al (2005) Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 237:395–400
16. Seemann MD, Seemann O, Luboldt W et al (2000) Differentiation of malignant from benign solitary pulmonary lesions using chest radiography, spiral CT and HRCT. Lung Cancer 29:105–124
17. Ponder BA (2001) Cancer genetics. Nature 411:336–341
18. Knudson AG (2002) Cancer genetics. Am J Med Genet 111:96–102
19. Peto J (2001) Cancer epidemiology in the last century and the next decade. Nature 411:390–395
20. Perera FP (2000) Molecular epidemiology: on the path to prevention? J Natl Cancer Inst 92:602–612
21. Hemminki K, Mutanen P (2001) Genetic epidemiology of multistage carcinogenesis. Mutat Res 473:11–21
22. Vazquez M, Carter D, Brambilla E et al (2009) Solitary and multiple resected adenocarcinomas after CT screening for lung cancer: histopathologic features and their prognostic implications. Lung Cancer 64:148–154
23. Varoli F, Vergani C, Caminiti R et al (2008) Management of solitary pulmonary nodule. Eur J Cardiothorac Surg 33:461–465
24. Wahidi MM, Govert JA, Goudar RK et al (2007) Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer? ACCP Evidence-based Clinical Practice Guidelines, 2nd edn. Chest 132(3 Suppl):94S–107S