Risk factors and prognostic nomogram for patients with second primary cancers after lung cancer using classical statistics and machine learning

lianxiang luo (luolianxiang321@gdmu.edu.cn)  
Guangdong Medical College Zhanjiang Campus: Guangdong Medical University  
https://orcid.org/0000-0002-3391-9713

Haowen Lin  
Guangdong Medical College Zhanjiang Campus: Guangdong Medical University

Baixin Lin  
Guangdong Medical College Zhanjiang Campus: Guangdong Medical University

Fangfang Huang  
Guangdong Medical College Zhanjiang Campus: Guangdong Medical University

Hui Luo  
Guangdong Medical College Zhanjiang Campus: Guangdong Medical University

Research Article

Keywords: Second primary cancer, Lung cancer, Risk factor, Nomogram, Survival

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1256731/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Background: Previous studies have revealed an increased risk of secondary primary cancers (SPC) after lung cancer. The prognostic prediction models for SPC patients after lung cancer are particularly needed to guide screening. Therefore, we study retrospectively analyzed the Surveillance, Epidemiology, and End Results (SEER) database using classical statistics and machine learning to explore the risk factors and construct a novel OS prediction nomogram for patients with SPC after lung cancer.

Methods: Data of patients with SPC after lung cancer, covering 2000 to 2016, were gathered from the Surveillance, Epidemiology, and End Results (SEER) database. The incidence of SPC after lung cancer was calculated by Standardized incidence ratios (SIRs). Cox proportional hazards regression, Kaplan-Meier methods and log-rank tests were conducted to identify the independent prognostic factors for predicting OS. These significant prognostic factors were used for the development of an OS prediction nomogram.

Results: Totally, 10487 SPC samples gathered from the SEER database were divided into training cohort and validation cohort (score construction and internal validation) at random. Multivariate Cox regression analysis showed that age, sex, race, grade, marital status, CT (chemotherapy), TNM stage, radiation, lung cancer surgery and SPC surgery were independent risk factors for OS in SPC patients. The prognostic nomogram we constructed was also for OS. The concordance index (C-index) in the training cohort and validation cohort were 0.715(95%CI:0.710-0.718) and 0.716(95%CI:0.712-0.720). Additionally, calibration curves and decision curve analysis (DCA) curves revealed that the nomogram has excellent clinical utility. In the random forest (RF) model, SPC site and T stage were the two most important variables, which was consistent with the nomogram. And, the random forest (RF) model had slightly better prediction performance than the Cox risk regression model (AUC: 0.748 vs. AUC: 0.719).

Conclusions: The prognosis characteristics of SPC following the lung cancer was systematically reviewed. The nomogram we constructed can provide survival predictions to assist clinicians in making individualized decisions for appropriate treatment.

Introduction

Lung cancer is a malignancy with both the highest incidence rate and mortality rate in the world\(^1\). With the promotions of screening programs and the development of treatment for lung cancer, the number of long-term survivors is gradually increasing\(^2,3\). However, the risk of SPC among lung cancer survivors is also on the increase, and their prognostic and survival analyses cannot be neglected.

According to recent studies, second primary lung cancers or other malignancy occur in 33% of patients with non-small cell lung carcinoma\(^4\). The second cancer of cancer survivors: New Malignancies Among Cancer Survivors: SEER Cancer Registries, 1973-2000 published by the National Cancer Institute (NCI), the probability of second cancer after the first cancer is higher than that of ordinary people. Taking lung cancer as an example, from 1973 to 2000, a total of 255083 patients with lung cancer were registered by...
SEER database, of which 11700 patients developed the second primary cancer. After adjusting the influencing factors, it was found that the incidence of second primary cancer was 5.7%, of which 29% was recurrent lung cancer. The data show that recurrence of second primary tumor in lung cancer patients within 20 years of diagnosis is very high probable. But the outcome and survival analyses of these SPCs are not fully understood. Therefore, the prognostic prediction models for SPC patients after lung cancer are particularly needed to guide screening.

Based on regression analysis, nomogram is created and plays an important role in this digital era. It can graphically represent the effect of each predictor variable on the outcome and provides a more intuitive and specific interpretation. Its efficiency has been demonstrated and has even become a new standard. At present, it is widely used to predict the recurrence or prognosis of various cancers such as prostate cancer, liver cancer, and many other diseases.

Random forest is a machine learning method of the regression tree technology, which achieves a high degree of prediction accuracy by using bootstrap aggregation and randomization of prediction variables. At present, many studies have shown that the random forest model is superior to the traditional model in many aspects. By combining the voting results of multiple decision trees to perform classification or regression, it has been widely used in the field of clinical research and bioinformatics. However, as a new model, random forest also has some disadvantages. For example, it is vulnerable to outliers. The Cox proportional hazard regression model is not only used to predict the survival data, but also can easily give the relationship between variables and survival outcome. Therefore, the random survival forest model should be combined with the traditional survival analysis, and can not completely replace the traditional survival analysis model. This combination may be more suitable for clinical research.

In this study, we aimed to investigate the general features affecting the prognosis of SPCs, and to construct a nomogram that predicts the 3-year and 5-year survival in SPC patients.

**Materials And Methods**

**Collection of patient data**

The case of SPCs was extracted from 18 population-based registries (2000–2016) in the Surveillance, Epidemiology, and End Results (SEER) database using SEER* Stat version 8.3.8. SEER program is the most authoritative and main source of cancer statistics aiming to reduce the cancer burden of American population. The tumor information in the database is unified and standardized through SEER* stat software (accession number: 15132-Nov2020), and is updated and released regularly.

Lung cancer diagnoses were defined using the SEER site recode International Disease for Oncology (ICD-O) and third edition site codes C34.0-C34.9. Patients who suffer from SPC after initial primary lung cancer (IPLC) were enrolled. The Warren criterion was used to identify SPC: histologically distinct from
primary cancer, with a latency of no less than 6 months to exclude errors caused by metastasis and recurrence\textsuperscript{13}.

In a total of 43707 patients with lung and bronchus disease diagnosed during 2000-2016 were downloaded from the SEER database. The inclusion criteria were as follows: 1. Patients with complete survival data and follow-up information. 2. Patients with “positive histology” (considered to have a pathologic diagnosis). 3. latency period between IPLC and SPC is more than 6 months.

The exclusion criteria were as follow: 1. patients with only autopsy or death certificate records. 2. the SPC cohort was identified as cancers with the same histology as the IPLC. 3. patients with unknown survival time.

In this study, in the case of SIR in SPC, the start of our follow-up was the date of diagnosis record by default in SEER database. Overall survival (OS) was calculated from the diagnosis date of the second primary cancer (latency period between IPLC and SPC is more than 6 months) to the date of the last follow-up or death in the SEER database.

**Patient cohorts**

Relative demographic and clinicopathological information was extracted including age, sex, grade, race, TNM stage (the most common tumor staging system in the world), stage (tumor metastasis stage), lymph node dissection of SPC, radiation, chemotherapy, site of SPC, histological type of SPC, surgery history of SPC and IPLC, marital status, follow-up time, and latency time between IPLC and SPC. The group of 60-65 years was a key age in lung cancer research. Studies had shown that the majority of lung cancer developed at the age of 65\textsuperscript{14}. Many references used this as a cut-off point for age groups \textsuperscript{15, 16}. In this study, we choose 60 as the cut-off point for age. In the SPC risk study, the study population was patients with lung cancer. In the survival analysis, the study population was patients with SPC after lung cancer. The purpose of the risk of SPC is to study the population in which SPC is prone to develop, so as to further eliciting our research on the prognosis of SPC in IPLC survivors. The SPC risk study and the study of overall survival in SPC were step wise and complement each other.

**SIR calculation**

The incidence of SPC in the cohort of patients with IPLC was compared with the expected incidence of lung malignancy in the general population using the SIR. SIR analyses were completed using the SEER*Stat program. The SIR was a calculation of the risk of developing a second cancer in patients with lung cancer by comparison with the incidence of cancers in the general population (in the same period as lung cancer patients). Each variable was individually selected before calculating and they were not distinguished by ethnicity. On study cohort selection, patients with the occurrence of a first lung malignancy primary registered in the SEER database were selected. The latency period between IPLC and SPC was more than 6 months to exclude errors caused by metastasis and recurrenceWarren S. Multiple primary malignant tumor. A surgery of the literature and statistical study [J]. Am J Cancer, 1932, 16..
According to clinical experience, the diagnosis of the second primary cancer in children was inconsistent with that in adults, hence patient less than 20 was excluded. At the same time, samples with unknown age variable in the cohort were also excluded.

**Prognosis related data analysis**

Samples were divided into training cohort (accounting for 70%, 7343) and validation cohort (accounting for 30%, 3144) by randomly using R package. Training cohort was used to establish the mathematical model after which validation cohort was used to test and verify the accuracy of the model. Significant prognostic factors were integrated to construct a nomogram. The internal validation was conducted respectively in the validation set.

Based on the Cox regression analysis, we chose the significant prognostic factors to evaluate the OS rate with Kaplan–Meier curves. The differences in survival were assessed using the log-rank test. A nomogram was developed from the training set to predict the 3-year survival and 5-year survival rates in SPC patients, using the rms, foreign and survival packages in R software. In order to assess the predictive accuracy of the nomogram model, the concordance index (C-index) was calculated by the Hmisc package in R software. The value of the C-index ranges from 0.5 to 1, where 0.5 indicates a random chance and 1 indicates that the model can correctly discriminate the outcome. Also, a calibration curve was drawn to compare the consistency between the predicted probability and the actual result. In addition, DCA, a new method for highlighting models of prediction with clinical net benefits was carried out to evaluate the clinical net benefit of the prognostic nomogram, and this new model was compared with the 8th edition AJCC TNM staging system.

**Random forest model construction**

Based on randomForest, caret and pROC packages of R software, we used the training set to establish a random forest (RF) model. With the help of out of bag data (OOB) error rate as a reference, we select nTree = 500 and mtry = 7 which had the relatively low OOB error rate. In order to evaluate which indicators had better prediction results for the model, 12 variables were sorted according to the decline of average Gini value. A higher value means that the variable was more important. Based on the prediction of the test set, the ROC curve of random forest model was drawn and the area under the curve was calculated. The prediction accuracy of the model could be judged by the AUC area value (0.5-1) under the subject ROC curve. The closer the AUC was to 1, the better the diagnostic effect. In addition, we also drew the ROC curve of Cox risk regression model to identify the advantages and disadvantages of prediction ability by comparing the AUC values of the two models.

**Result**

**Analysis of patient data**
Ultimately, a total of 10487 patients who presented with a second primary tumor after being diagnosed with lung cancer were identified. The detailed patient selection process was shown in Figure 1. The results of patient samples revealed that the median age of SPCs patient was 72 and the median survival time was 20 months. Features of second primary cancer (SPC) after lung cancer (IPLC) had shown in Figure 2. Among them, Adenomas and adenocarcinomas, Ductal, lobular and medullary neoplasms, Transitional cell papillomas and carcinomas, Squamous cell neoplasms, Hodgkin and non-Hodgkin lymphoma were the top 5 histological types of SPC in turn (Figure 2, A). Urinary System, Male Genital System, Colon and Rectum, Breast, Lymphoma were the top 5 items in tumor site (Figure 2, B). The details were showed in Supplementary table 1 and Supplementary table 2.

**SIR analysis**

The SIR of SPC among IPLC survivors were listed in Table 1. Univariate and bivariate comparison were used to calculate SIR, such as the longitudinal comparison between women and women after surgery. A comprehensive conclusion can be drawn from the horizontal comparison of female and male. To make the results more intuitive, lines with observe = 0 were not shown in the table. A SIR greater than 1 suggests that individuals who have lung cancer were at increased risk of developing another type of cancer as compared to the general population. *P*-value ≤ 0.05 proved statistically different. In the SIR analysis, we performed the relevant calculations for fields that are clinically common. The results showed that age, grade, CT, laterality, LNR, radiation, sex, stage and marital status were highly correlated with the occurrence of SPC.
Table 1
SIR of SPC after lung cancer.

| Characteristic                  | Observe | O/E |
|--------------------------------|---------|-----|
| **Age**                        |         |     |
| 20-24 years                    | 5       | 5.81# |
| 25-29 years                    | 9       | 4.03# |
| 30-34 years                    | 22      | 4.02# |
| 35-39 years                    | 79      | 4.85# |
| 40-44 years                    | 217     | 3.52# |
| 45-49 years                    | 622     | 3.11# |
| 50-54 years                    | 1,416   | 2.66# |
| 55-59 years                    | 2,561   | 2.27# |
| 60-64 years                    | 3,883   | 2.07# |
| 65-69 years                    | 4,696   | 1.79# |
| 70-74 years                    | 4,269   | 1.63# |
| 75-79 years                    | 3,038   | 1.48# |
| 80-84 years                    | 1,363   | 1.30# |
| 85+ years                      | 347     | 1.05 |
| **CT**                         |         |     |
| No/Unknown                     | 15,320  | 1.87# |
| Yes                            | 7,210   | 1.67# |
| **Grade**                      |         |     |
| Well differentiated; Grade I   | 2,595   | 1.83# |
| Moderately differentiated; Grade II | 7,221 | 1.97# |
| Poorly differentiated; Grade III | 7,122  | 1.89# |
| Undifferentiated; anaplastic; Grade IV | 823 | 1.66# |
| Unknown                        | 4,769   | 1.52# |
| **Laterality**                 |         |     |
| Not a paired site              | 11      | 2.27# |
| Right - origin of primary      | 12,899  | 1.80# |
| Characteristic                                                                 | Observe | O/E  |
|-------------------------------------------------------------------------------|---------|------|
| Left - origin of primary                                                      | 9,371   | 1.84#|
| Only one side - side unspecified                                              | 36      | 1.70#|
| Bilateral, single primary                                                     | 56      | 1.12 |
| Paired site, but no information concerning laterality                         | 157     | 1.31#|
| **LNR**                                                                       |         |      |
| None                                                                          | 5,181   | 1.46#|
| Biopsy or aspiration of regional lymph node, NOS                              | 425     | 1.75#|
| Sentinel lymph node biopsy                                                    | 28      | 2.80#|
| Number of regional lymph nodes removed unknown                                | 600     | 1.86#|
| 1 to 3 regional lymph nodes removed                                           | 2,386   | 2.11#|
| 4 or more regional lymph nodes removed                                        | 8,759   | 2.07#|
| Sentinel node biopsy and lym nd removed same/unstated time                    | 51      | 2.38#|
| Sentinel node biopsy and lym nd removed different times                       | 21      | 2.08#|
| Unknown or not applicable                                                     | 84      | 1.45#|
| Blank(s)                                                                      | 4,995   | 1.72#|
| **Radiation**                                                                 |         |      |
| None/Unknown                                                                  | 16,535  | 1.86#|
| Beam radiation                                                                | 5,676   | 1.68#|
| Radioactive implants (includes brachytherapy) (1988+)                         | 35      | 1.90#|
| Combination of beam with implants or isotopes                                 | 27      | 1.92#|
| Radiation, NOS method or source not specified                                | 85      | 1.53#|
| Refused (1988+)                                                               | 68      | 1.21 |
| Recommended, unknown if administered                                          | 104     | 1.43#|
| **Sex**                                                                       |         |      |
| Male and female                                                               | 22,530  | 1.80#|
| Male                                                                          | 11,631  | 1.66#|
| Female                                                                        | 10,899  | 1.99#|
| **Stage**                                                                     |         |      |
| Characteristic                  | Observe | O/E |
|--------------------------------|---------|-----|
| Localized                      | 11,138  | 1.99# |
| Regional                       | 8,056   | 1.85# |
| Distant                        | 2,904   | 1.30# |
| Unknown/unstaged               | 432     | 1.48# |

**Marital status**

| Marital status                        | Observe | O/E |
|---------------------------------------|---------|-----|
| Single (never married)                | 2,192   | 1.88# |
| Married (including common law)        | 13,446  | 1.80# |
| Separated                             | 211     | 1.97# |
| Divorced                              | 2,645   | 2.04# |
| Widowed                               | 3,211   | 1.63# |
| Unmarried or Domestic Partner         | 12      | 1.83 |
| Unknown                               | 813     | 1.73# |

**Marital status & Sex**

**Male and female**

| Marital status                        | Observe | O/E |
|---------------------------------------|---------|-----|
| Single (never married)                | 2,192   | 1.88# |
| Married (including common law)        | 13,446  | 1.80# |
| Separated                             | 211     | 1.97# |
| Divorced                              | 2,645   | 2.04# |
| Widowed                               | 3,211   | 1.63# |
| Unmarried or Domestic Partner         | 12      | 1.83 |
| Unknown                               | 813     | 1.73# |

**Male**

| Marital status                        | Observe | O/E |
|---------------------------------------|---------|-----|
| Single (never married)                | 1,127   | 1.73# |
| Married (including common law)        | 8,146   | 1.66# |
| Separated                             | 118     | 2.00# |
| Divorced                              | 1,116   | 1.88# |
| Widowed                               | 760     | 1.39# |
| Unmarried or Domestic Partner         | 6       | 1.55 |
| Characteristic                                      | Observe | O/E |
|----------------------------------------------------|---------|-----|
| Unknown                                            | 358     | 1.46\# |
| **Female**                                         |         |     |
| Single (never married)                             | 1,065   | 2.08\# |
| Married (including common law)                     | 5,300   | 2.05\# |
| Separated                                          | 93      | 1.94\# |
| Divorced                                           | 1,529   | 2.18\# |
| Widowed                                            | 2,451   | 1.73\# |
| Unmarried or Domestic Partner                      | 6       | 2.23 |
| Unknown                                            | 455     | 2.03\# |
| **Sex & Age**                                      |         |     |
| **Male and female**                                |         |     |
| 20-24 years                                        | 5       | 5.81\# |
| 25-29 years                                        | 9       | 4.03\# |
| 30-34 years                                        | 22      | 4.02\# |
| 35-39 years                                        | 79      | 4.85\# |
| 40-44 years                                        | 217     | 3.52\# |
| 45-49 years                                        | 622     | 3.11\# |
| 50-54 years                                        | 1,416   | 2.66\# |
| 55-59 years                                        | 2,561   | 2.27\# |
| 60-64 years                                        | 3,883   | 2.07\# |
| 65-69 years                                        | 4,696   | 1.79\# |
| 70-74 years                                        | 4,269   | 1.63\# |
| 75-79 years                                        | 3,038   | 1.48\# |
| 80-84 years                                        | 1,363   | 1.30\# |
| 85+ years                                          | 347     | 1.05 |
| **Male**                                           |         |     |
| 20-24 years                                        | 3       | 9.70\# |
| 25-29 years                                        | 4       | 5.42\# |
| Characteristic | Observe | O/E |
|---------------|---------|-----|
| 30-34 years   | 9       | 5.64#|
| 35-39 years   | 25      | 5.00#|
| 40-44 years   | 72      | 3.36#|
| 45-49 years   | 245     | 3.00#|
| 50-54 years   | 701     | 2.66#|
| 55-59 years   | 1,299   | 2.06#|
| 60-64 years   | 2,117   | 1.92#|
| 65-69 years   | 2,451   | 1.62#|
| 70-74 years   | 2,280   | 1.51#|
| 75-79 years   | 1,556   | 1.36#|
| 80-84 years   | 684     | 1.23#|
| 85+ years     | 184     | 1.07 |

**Female**

| Characteristic | Observe | O/E |
|---------------|---------|-----|
| 20-24 years   | 2       | 3.63 |
| 25-29 years   | 5       | 3.35#|
| 30-34 years   | 13      | 3.35#|
| 35-39 years   | 54      | 4.78#|
| 40-44 years   | 145     | 3.60#|
| 45-49 years   | 377     | 3.18#|
| 50-54 years   | 715     | 2.66#|
| 55-59 years   | 1,262   | 2.54#|
| 60-64 years   | 1,766   | 2.28#|
| 65-69 years   | 2,245   | 2.04#|
| 70-74 years   | 1,989   | 1.79#|
| 75-79 years   | 1,482   | 1.62#|
| 80-84 years   | 679     | 1.39#|
| 85+ years     | 163     | 1.03 |

Abbreviations: SPC: second primary cancer; LNR: lymph node removed; CT: chemotherapy; #: P<0.05
Characteristics of SPC

The result of prognostic factors for the OS of SPC were all listed in Table 2. Univariate Cox regression analysis revealed that age, sex, race, CT, grade, LNR, marital status, TNM stage, radiation, surg history of IPLC and surg history of SPC were independent prognostic factors the OS of SPC. As for multivariate Cox regression analysis, age, race, sex, marital status, CT, TNM stage, radiation, grade, lung cancer surgery and SPC surgery were associated with for the OS of SPC.

Table2. Univariate and multivariate Cox regression analysis.
| Characteristics | Univariate analysis | Multivariate analysis |
|-----------------|--------------------|----------------------|
|                 | HR  | 95%CI    | P       | HR  | 95%CI    | P      |
| Age             |     |          |         |     |          |        |
| <60             | Reference |           |         | Reference |           |        |
| ≥60             | 1.36 | 1.24 - 1.49 | 9.24e-11 | 1.40 | 1.28 - 1.54 | 5.90e-12 |
| Race            |     |          |         |     |          |        |
| White           | Reference |           |         | Reference |           |        |
| Black           | 1.03 | 0.93 - 1.13 | 0.568   | 0.89 | 0.81 - 0.98 | 0.020  |
| Other           | 0.81 | 0.70 - 0.94 | 0.005   | 0.72 | 0.70 - 0.95 | 0.008  |
| Sex             |     |          |         |     |          |        |
| Male            | Reference |           |         | Reference |           |        |
| Female          | 0.86 | 0.81 - 0.91 | 2.43e-07 | 0.87 | 0.81-0.93 | 2.00e-05 |
| Marital status  |     |          |         |     |          |        |
| Single          | Reference |           |         | Reference |           |        |
| Married         | 0.81 | 0.77-0.87 | 3.02e-11 | 0.81 | 0.76-0.86 | 6.75e-11 |
| Unknown         | 0.82 | 0.73 – 0.92 | 0.000452 | 0.72 | 0.64-0.81 | 1.85e-08 |
| CT              |     |          |         |     |          |        |
| No              | Reference |           |         | Reference |           |        |
| Yes             | 1.36 | 1.27 - 1.45 | <2e-16  | 1.2 | 1.13 - 1.31 | 1.99e-07 |
| LNR             |     |          |         |     |          |        |
| No              | Reference |           |         | Reference |           |        |
| Yes             | 0.79 | 0.72 - 0.86 | 3.36e-08 | 0.96 | 0.87-1.06 | 0.43   |
| Unknown         | 0.94 | 0.88 - 1 | 0.044 | 1.05 | 0.97-1.13 | 0.22   |
| Surgery history of IPLC | | | | | | |
| No              | Reference |           |         | Reference |           |        |
| Yes             | 0.47 | 0.44 - 0.5 | <2e-16  | 0.49 | 0.46-0.53 | < 2e-16 |
| Surgery history of SPC | | | | | | |
| No              | Reference |           |         | Reference |           |        |
| Yes             | 0.6  | 0.56 - 0.63 | <2e-16  | 0.61 | 0.57 - 0.66 | < 2e-16 |
| Radiation | Yes | 0.77 | 0.71 - 0.82 | 2.57e-13 | 0.71 | 0.65 - 0.77 | < 2e-16 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Grade I | Reference | Reference | Grade II | 1.25 | 1.10-1.41 | 0.000433 | 1.11 | 0.98-1.26 | 0.11 |
| Grade III | 1.59 | 1.40 - 1.80 | 8.16e-13 | 1.11.26 | 1.10-1.43 | 0.0005 |
| Grade IV | 2.32 | 1.96-2.74 | < 2e-16 | 1.96 | 1.65-2.32 | 6.99e-15 |
| Other | 2.03 | 1.807-2.28 | < 2e-16 | 1.52 | 1.34-1.71 | 1.53e-11 |

| Stage_T | T0 | Reference | Reference | T1 | 0.11 | 0.05 – 0.24 | 1.07e-08 | 0.40 | 0.19-0.87 | 0.20 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stage_M | M0 | Reference | Reference | M1 | 3.70 | 3.23 - 4.23 | < 2e-16 | 2.18 | 1.88-2.52 | < 2e-16 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stage_N | N0 | Reference | Reference | N1 | 1.61 | 1.39 - 1.87 | 1.56e-10 | 1.23 | 1.05-1.44 | 0.011 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Abbreviations: SPC second primary cancer; LNR lymph node removed; CT chemotherapy; Grade I Well differentiated; Grade II Moderately differentiated; Grade III Poorly differentiated; Grade IV Undifferentiated; anaplastic; “Other” of the “Grade” includes “unknown, T cell, B cell, null cell”. “Other” of the “Stage_T” includes “Tis, TX, Ta, NA, Blank”. "Other" of the “Stage_N” includes “NX, NA, Blank".
Kaplan–Meier analysis for prognostic factors.

Based on the above study, twelve variables were selected for KM analysis. Figure 3A-L depict the Kaplan-Meier survival curves for overall survival (OS) of second primary cancer (SPC) after lung cancer (LC) based on risk score in the training set. Because of the paucity of patients with definite M stage in the data, it was not analyzed. Similarly, race and marital status were also excluded. The log-rank p-value for these prognostic factors in the training set were all<0.01, which suggested that these characteristics significantly affected patient survival. Patients who received treatment like lung cancer surgery, SPC surgery or radiation tended to have increased survival. But chemotherapy was a risk factor for death. Also, Ductal, lobular and medullary neoplasms had more OS benefits based on SPC histology. SPC in male genital system and breast appeared to have longer OS than another site.

Analysis of Random Forest prediction model

The RF model was utilized for variable importance ranking, the two most important variables were SPC site and T stage (Figure 4A). When the number of decision trees was 500, the three lines were in a smooth and insignificantly fluctuating state, indicating that the random forest was able to well discriminate the survival status of SPC patients. The red line was the error rate of 1, the green line was the error rate of 0, and the black line is the OOB error rate (Figure 4B). The area under the curve (AUC) of this model was 0.748 (95% Cl: 0.667-0.828), which indicated a certain accuracy between probability and observation in the RF model (Fig. 4C). The ROC curves of Cox risk regression model were shown in supplementary Figure 1. We found that the RF model had better specificity as well as sensitivity and superior predictive performance than the Cox risk regression model (AUC: 0.748 vs. AUC: 0.719).

Construction of prognostic nomogram

Based on the training set, the nomogram predicting 3-year and 5-year probabilities for OS were established using statistically significant variables from Cox analysis and RF model, as shown in Figure 5. Twelve clinical indicators were incorporated in the nomogram, including surgery of LC, surgery of SPC, site, histology, sex, age, chemotherapy, radiation, grade, LNR, stage_T and stage_N. These factors were used to accumulate the corresponding scores and overall score to determine the probability of survival for SPC after the I PLC. The nomogram revealed that site of SPC and stage_T had the greatest influence on patient prognosis, followed by histology of SPC and surgery of SPC. The top 5 significant features were also largely consistent with the main prognostic impact factors shown by the RF model. In particular, the nomogram was scored by univariate comparison. In other words, when SPC was found in the breast, the patient’s gender was confirmed to be female and had nothing to do with the sex above. Therefore, our nomogram is applicable to both male and female patients without conflict.

Calibration and validation of the nomogram
The calibration curves were used to evaluate the nomogram with both training set and validation set in Figure 6. The discriminative ability and predictive accuracy of the nomogram were evaluated by concordance index (C-index), calibration curve and DCA curves. The nomogram indicated good predictive performance with the C-index of 0.715 (95%CI:0.710-0.718) in the training cohort (Figure 6A-B), and 0.716 (95%CI:0.712-0.720) in the validation cohort (Figure 6C-D). The abscissa of the calibration curve was the prediction probability, and 0 to 1 indicated that the probability of an event is 0 to 100%. The ordinate was the actual probability: the actual event rate of the patient. The red line was the fitting line, indicating the actual value corresponding to the predicted value. The accuracy of the model was intuitively demonstrated by the coincidence degree of the fitting line. If the predicted value was equal to the actual value, the red line completely coincides with the reference line (blue line). If the predicted value was greater than the actual value, the risk was overestimated and the red line was above the blue line. If the predicted value was less than the actual value, the risk was underestimated and the red line was above the blue line.

The calibration curves displayed that a favorable consistency between the nomogram predictions as well as the actual outcomes of the 3-year (Figure 6A, C) and 5-year (Figure 6B, D) survival in both training cohort and validation cohort. These indicated that our nomogram was judged to have good discrimination and prediction abilities in the internal validation.

Additionally, DCA was implemented in the training cohort and validation cohort to ascertain the clinical usefulness of the nomogram (Figure 7). The results showed that both 3-year survival (Figure 7A) and 5-year survival(Figure 7B) prediction in our model performed significantly better than the 8th edition AJCC TNM staging system, with the higher C-index and better net benefits in predicting the OS of SPC patients.

**Discussion**

The survival rate of lung cancer has been greatly improved due to the development of medical treatment\(^2\). The five-year survival rate of lung cancer patients in the United States has increased from 17.2% ten years ago to 21.7% now, according to the American Lung Association, a report released on November 12 called the state of lung cancer\(^18\). But at the same time, cancer survivors have a significantly increased risk of developing new cancers\(^19\). Among them, female are more likely to suffer from second primary cancers\(^20\). Patients with pneumonia have an increased risk of secondary primary malignancy\(^21\). Mortality from second primary malignancies is very high, even exceeding that of patients with a single cancer\(^22\). The treatment of cancer is constantly improving and changing. But from 2000 to now, surgery, chemotherapy and radiation remain standard treatment for many kinds of cancer, with current immune-targeting therapies applying to only a small subset of patients\(^23\). Nomogram created by regression analysis has been widely employed to predict the prognosis of diverse cancers because of its simplicity, intuitiveness, and practicality\(^24,25\). Hence, the study of clinical characteristics and overall survival nomogram of SPCs after lung cancer is more than important. In this study, we aimed to investigate the general features affecting the prognosis of SPCs, and to construct a nomogram that predicts the 3-year
and 5-year survival in SPC patients, since there is no such survival predicting model for SPC after lung cancer presented as nomogram.

In our study, female was more likely to suffer from second primary cancers, which was consistent with previous studies\(^\text{20}\). The SIR of male in 20-24 years (5.81) was higher than that of male in other age, followed by and 30-34 years (5.64). The SIR of female in 35-39 years (4.78) was higher than that of female in other age. However, due to the small sample size of patients younger than 50 years old, we cannot draw a definite conclusion whether this age group has the highest risk. But it can be seen that patients with SPC after IPLC tend to be younger. Particularly, there is little significant difference between male and female only in sex analysis. But when age was added, the SIR of female (age>55) was much higher than that of male (age>55). The SIR of other different characteristics is relatively average and all greater than 1, suggesting the increased risk of developing another type of cancer as compared to the general population.

As for OS, hazard ratio (HR) was used to compare the risk of death. In the multivariate Cox regression analysis of age, HR of age \(\geq 60\) to age \(\leq 60\) is 1.36, which indicated that people older than 60 have higher risk. Also, black people were more at risk than white people. The risk was greater for male than for female. In addition, LNR, radiation, surgery of the initial primary lung cancer (IPLC) and surgery of the second primary cancer (SPC) all reduced the risk of death and were all statistically significant. But chemotherapy would increase the risk of death. Besides, in Stage_T, we used T0 (carcinoma in situ) as a reference and found that death risk uniformly progressed from “T1” to “T4”, that is, the larger the tumor, the higher the risk. In Stage_N, the risk of lung cancer patients without regional lymph node metastasis was greatly reduced, which was common with the general studies. Noteworthy, “Other” was significant for OS in TMN stage. It was suggested that the definition criteria for primary tumor in TMN staging system were imperfect and should also be modified based on the analysis of more clinical data as well as clinical experience.

Although previous clinical data analysis reports have discussed the second primary cancer following lung cancer\(^\text{26,27}\), they have not constructed a nomogram that can be used to predict 3-year and 5-year survival rate with the clinical features. Recently, a study reported that the most common cancer of SPCs after lung cancer were prostate, breast, gastrointestinal tract\(^\text{28}\), and another study reported that were head and neck cancer, bone and soft tissue cancer, genitourinary cancer, and thyroid cancer\(^\text{13}\). Our study was generally consistent with them. In addition, the development of cancer in the urinary system following lung cancer may be due to share similar etiologic features, such as common carcinogenic pathway and chronic inflammatory substances\(^\text{29,30}\). Collectively, the prone sites as mentioned should be cautiously monitored.

From the perspective of the result of C-index and calibration curves, our model performed well in the evaluation. However, there has been reported that C-index is not clinically relevant and lack clinical practicability recently\(^\text{31}\). Therefore, the emerging DCA curve was used to help doctors judge the clinical net benefit of predictive models and make better clinical decisions\(^\text{32,33}\). To further assess the clinical utility of our model, we compared it with the 8th edition AJCC TNM staging system commonly used in
clinical evaluation using DCA curves, and the result showed that the predictive of our nomogram is superior. More importantly, the important variables we screened were verified by RF model. In other word, our nomogram was attested to the excellent sensitivity and good clinical value.

The retrospective study summarized the clinically significant indicators of SPC after lung cancer on the overall survival, and then constructed a clinical prognostic nomogram. It has not been reported to use nomogram to study SPC after IPLC. Whether studying bone metastasis of lung cancer or predicting death rate among non-small cell lung cancer (NSCLC) patients with surgery, the verification and the accuracy of our model were similar to them, which proved that our model was reliable in establishment and verification. The important difference is that we added Decision Curve Analysis (DCA) to assess the potential clinical utility of predictive nomogram and used RF model to test the importance of variables. And this is where we are better than them in validating the model. Random forest models have been widely used in clinical research and are also the model of choice among numerous machine learning.

The Cox proportional hazard regression model is not only used to predict the survival data, but also can easily give the relationship between variables and survival outcome. Therefore, the random survival forest model should be combined with the traditional survival analysis to help clinical decision-making.

Nevertheless, there are several limitations present in our study. First, the data used in the study was extracted from the SEER database, which may carry inherent biases. Second, the SEER database lacks information regarding target therapies and immunotherapy, making our model unable to consider all the influence factors. Third, even though we have verified the reliability of this model with c-index, calibration curves, DCA curves and RF model, it is still limited, and needs to be further validated in prospective clinical trials.

In conclusion, the nomogram we constructed can be used to effectively evaluate the 3-year and 5-year survival rate for the second primary cancer after lung cancer. Our study can help clinicians to make clinical decisions, and guide follow-up management strategies for patients with lung cancer.

**Declarations**

**Funding**

This project was supported by the Basic and Applied Basic Research Program of Guangdong Province (2019A1515110201); Program of Department of Natural Resources of Guangdong Province ([2021]53); Discipline Construction Project of Guangdong Medical University (4SG21004G).

**Competing interests**

The authors declare that they have no competing interests.

**Availability of data and materials**
The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Authorship contributions**

LX Luo conceived the idea; LX Luo, HW Lin, BX Lin and FF Huang contributed to the acquisition, analysis, and interpretation of data. LX Luo, HW Lin wrote the manuscript; LX Luo and H Luo reviewed the paper and provided comments, and all authors reviewed the manuscript.

**References**

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2020. *CA: A Cancer Journal for Clinicians*. 2020;70:7-30.
2. S W. Multiple primary malignant tumors. A survey of the literature and a statistical study. *Am J Cancer*. 1932;16.
3. Carroll NM, Burnett-Hartman AN, Joyce CA, et al. Real-world Clinical Implementation of Lung Cancer Screening-Evaluating Processes to Improve Screening Guidelines-Concordance. *J Gen Intern Med*. 2020;35:1143-1152.
4. Fink-Neuboeck N, Lindenmann J, Porubsky C, et al. Hazards of Recurrence, Second Primary, or Other Tumor at Ten Years After Surgery for Non-Small-Cell Lung Cancer. Clin Lung Cancer 2020;21:333–40.
5. Jr.1 RECDMFERLAGRDGHBKEMATJFF. New Malignancies Among Cancer Survivors:SEER Cancer Registries, 1973-2000. 2006.
6. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol*. 2008;26:1364-1370.
7. Park SY, Nomogram: An analogue tool to deliver digital knowledge. *J Thorac Cardiovasc Surg*. 2018;155:1793.
8. McEvoy SH, Raeside MC, Chaim J, Ehdaie B, Akin O. Preoperative Prostate MRI: A Road Map for Surgery. *AJR Am J Roentgenol*. 2018;211:383-391.
9. Pan YX, Chen JC, Fang AP, et al. A nomogram predicting the recurrence of hepatocellular carcinoma in patients after laparoscopic hepatectomy. *Cancer Commun (Lond)*. 2019;39:55.
10. Rigatti SJ Random Forest. *Journal of insurance medicine (New York, N.Y.)*. 2017;47:31-39.
11. Liu WC, Li ZQ, Luo ZW, Liao WJ, Liu ZL, Liu JM. Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer. *Cancer Med*. 2021;10:2802-2811.
12. Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. *Bioinformatics (Oxford, England)*. 2015;31:i197-205.
13. Vincent Yi-Fong Su T-JC, Liu Yu-WenHChia-Jen, Chen Sang-HueY,Yuh-Min, Tzeon-Jye Chiou, Teh-Ying Chou, Liu J-H, &. Risk of Second Primary Malignancies in Lung Cancer Survivors – The Influence of Different Treatments. *Targ Oncol*. 2016.
14. Gore E, Movsas B, Santana-Davila R, Langer C. Evaluation and management of elderly patients with lung cancer. Seminars in radiation oncology. 2012;22:304-310.
15. Yan B, Su BB, Bai DS, et al. A practical nomogram and risk stratification system predicting the cancer-specific survival for patients with early hepatocellular carcinoma. Cancer Med. 2021;10:496-506.
16. Park S, Park IK, Kim ER, et al. Current Trends of Lung Cancer Surgery and Demographic and Social Factors Related to Changes in the Trends of Lung Cancer Surgery: An Analysis of the National Database from 2010 to 2014. Cancer research and treatment. 2017;49:330-337.
17. Streiner DL, Cairney J. What's under the ROC? An introduction to receiver operating characteristics curves. Canadian journal of psychiatry. Revue canadienne de psychiatrie. 2007;52:121-128.
18. Editorial. Lung cancer: some progress, but still a lot more to do. Editorial. 2019.
19. Vogt ASS, Heinimann K, et al. Multiple primary tumours: challenges and approaches, a review. ESMO. 2017;2:e000172.
20. Smita Bhatia CS. second cancers in survivors of childhood cancer. 2002;2.
21. Wu X, Zhang X, Tao L, Chen P. Risk of second primary malignancy in adults with pulmonary high-grade neuroendocrine carcinoma (HGNEC). BMC Cancer. 2020;20:719.
22. Donin N, Filson C, Drakaki A, et al. Risk of second primary malignancies among cancer survivors in the United States, 1992 through 2008. Cancer. 2016;122:3075-3086.
23. Chen B, Alvarado DM, Iticovici M, et al. Interferon-Induced IDO1 Mediates Radiation Resistance and Is a Therapeutic Target in Colorectal Cancer. Cancer immunology research. 2020;8:451-464.
24. Zhang Y, Hong YK, Zhuang DW, He XJ, Lin ME. Bladder cancer survival nomogram: Development and validation of a prediction tool, using the SEER and TCGA databases. Medicine. 2019;98:e17725.
25. Feng SS, Li HB, Fan F, et al. Clinical characteristics and disease-specific prognostic nomogram for primary gliosarcoma: a SEER population-based analysis. Scientific reports. 2019;9:10744.
26. Barclay ME, Lyratzopoulos G, Walter FM, Jefferies S, Peake MD, Rintoul RC. Incidence of second and higher order smoking-related primary cancers following lung cancer: a population-based cohort study. Thorax. 2019;74:466-472.
27. Chuang SC, Scelo G, Lee YC, et al. Risks of second primary cancer among patients with major histological types of lung cancers in both men and women. Br J Cancer. 2010;102:1190-1195.
28. Faehling M, Schwenk B, Kramberg S, et al. Second malignancy in non-small cell lung cancer (NSCLC): prevalence and overall survival (OS) in routine clinical practice. J Cancer Res Clin Oncol. 2018;144:2059-2066.
29. Sanguedolce F, Loizzi D, Sollitto F, et al. Bladder Metastases from Lung Cancer: Clinical and Pathological Implications: A Systematic Review. Oncology. 2017;92:125-134.
30. Singh A, Kinoshita Y, Rovito PM Jr, et al. Higher than expected association of clinical prostate and bladder cancers. J Urol. 2008;179:S2-5.
31. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565-574.

32. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol*. 2018;74:796-804.

33. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.

34. Zheng XQ, Huang JF, Lin JL, et al. Incidence, prognostic factors, and a nomogram of lung cancer with bone metastasis at initial diagnosis: a population-based study. *Translational lung cancer research*. 2019;8:367-379.

35. Mogensen UB, Ishwaran H, Gerds TA. Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of statistical software*. 2012;50:1-23.

**Figures**

**Figure 1**

The overall framework of this study.

**Figure 2**

Features of second primary cancer (SPC) after lung cancer (IPLC). (A) Sites of SPCs that top of the list. (B) Histology types of SPCs that top of the list.

**Figure 3**

Kaplan–Meier analysis for overall survival (OS) of second primary cancer (SPC) after lung cancer (LC) based on risk score in the training set (p < 0.001). (A) chemotherapy (p < 0.001), (B) sex (p < 0.001), (C) age (p < 0.001) (D) SPC surgery (p < 0.001) (E) radiation (p < 0.001) (F) LC surgery, (G) LNR (p < 0.001) (H) stage_N (p < 0.001), (I) stage_T (p < 0.001), (J) grade (p < 0.001), (K) site, (L) histology.

**Figure 4**

Screening of important variables. (A) Rank order of significance of SPC related variables. (B) Dynamic relationship of prediction error to the number of decision trees. (C) ROC performance of the prediction
model at increasing number of features.

**Figure 5**

Nomogram to predict 3-year survival and 5-year survival for second primary cancer (SPC) patients.

**Figure 6**

The calibration curve to evaluate the 3-year (A) and 5-year (B) survival for second primary cancer (SPC) patients in training set. The calibration curve to evaluate the 3-year (C) and 5-year (D) survival for SPC patients in the validation set.

Nomogram - the predicted overall survival (OS) is plotted on the x-axis, and the actual OS is plotted on the y-axis. The imaginary line represents a perfect calibration model in which the predicted probability is the same as the actual survival result.

Standard deviation of our calibration curve: Training set: 3-years (min: 0.01331, median :0.02224, max.:0.04126); 5-years (min: 0.01907, median :0.03238, max.:0.05966). Validation set: 3-years (min: 0.02050, median: 0.03491, max:0.06566); 5-years (min: 0.03087, median :0.05132, max. :0.08850).

**Figure 7**

Decision curve analysis for 3-year(A) and 5-year(B) OS comparing the nomogram with the 8th edition AJCC TNM staging system in the training cohort. Decision curve analysis for 3-year (A) and 5-year (B) OS comparing the nomogram with the 8th edition AJCC TNM staging system in the validation cohort.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfigure1.png
- supplementarytable1.docx
- supplementarytable2.docx