# Predict multicategory causes of death in lung cancer patients using clinicopathologic factors

Fei Deng[1], Haijun Zhou[2], Yong Lin[3,4], John Heim[5], Lanlan Shen[6], Yuan Li[7,8]*, Lanjing Zhang[3,5,9, 10*]

[1] School of Electrical and Electronic Engineering, Shanghai Institute of Technology, China.

[2]Department of Pathology and Genomic Medicine, Houston Methodist Hospital, Houston, Texas.

[3]Rutgers Cancer Institute of New Jersey, Rutgers, New Brunswick, New Jersey.

[4]Department of Biostatistics, Rutgers School of Public Health, Piscataway, New Jersey.

[5]Princeton Medical Center, Plainsboro, New Jersey.

[6]Department of Pediatrics, Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, Texas.

[7]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China.

[8]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

[9]Department of Biological Sciences, Rutgers University, Newark, New Jersey.

[10]Department of Chemical Biology, Rutgers Ernest Mario School of Pharmacy, Rutgers University, Piscataway, Newark.

*YL and LZ equally supervised the works.

Correspondence: Lanjing Zhang, MD, Email: lanjing.zhang@rutgers.edu or Yuan Li, MD, Email: lumoxuan2009@163.com

## Abstract

**Background**: Random forest model is a recently developed machine-learning algorithm, and superior to other machine learning and regression models for its classification function and better accuracy. But it is rarely used for predicting causes of death in lung cancer patients. On the other hand, specific causes of death in lung cancer patients are poorly classified or predicted, largely due to its categorical nature (versus binary death/survival).

**Methods**: We therefore tuned and employed a random forest algorithm (Stata, version 15) to classify and predict specific causes of death in lung cancer patients, using the surveillance, epidemiology and end results-18 and several clinicopathological factors. The lung cancer diagnosed during 2004 were included for the completeness in their follow-up and death causes. The patients were randomly divided into training and validation sets (1:1 match). We also compared the accuracies of the final random forest and multinomial regression models.

**Results**: We identified and randomly selected 40,000 lung cancers for the analyses, including 20,000 cases for either set. The causes of death were, in descending ranking order, were lung cancer (72.45 %), other causes or alive (14.38%), non-lung cancer (6.87%), cardiovascular disease (5.35%), and infection (0.95%). We found more 250 iterations and the 10 variables produced the best prediction, whose best accuracy was 69.8% (error-rate 30.2%). The final random forest model with 300 iterations and 10 variables reached an accuracy higher than that of multinomial regression model (69.8% vs 64.6%). The top-10 most important factors in the random-forest model were sex, chemotherapy status, age (65+ vs <65 years), radiotherapy status, nodal status, T category, histology type and laterality, which were also independently associated with 5-category causes of death.

**Conclusion**: We optimized a random forest model of machine learning to predict the specific cause of death in lung cancer patients using a set of clinicopathologic factors. The model also appears more accurate than multinomial regression model.

# Introduction

Lung cancer is one of the leading causes of death in the U.S. and the world [1]. Basic research has revealed new insights into development and progression of lung cancer [2]. Clinical works also identified the factors associated with the survival outcomes of lung cancer, particularly non-small cell lung cancer, including histology and oncogenic drivers [3]. However, the survival outcomes of the lung cancers are mostly binary, which were either alive versus death or alive versus progression/death. Few studies on lung cancer patients were focused on the specific deaths due to non-lung cancer[4], cardiovascular diseases[5, 6], infection or other causes[4, 7]. Even fewer works investigated the multicategory causes of death (COD).

Random forest (RF) model is a recently developed machine-learning algorithm, and superior to other machine learning and regression models for its classification function and better accuracy [8]. But it is rarely used for predicting causes of death in lung cancer patients, while we have shown a better prediction accuracy in prostate cancers [8]. On the other hand, specific causes of death in lung cancer patients are poorly classified or predicted, largely due to its categorical nature (versus binary death/survival). Therefore, this study was aimed to understand the factors associated with multicategory COD in lung cancer patients using RF and multinomial regression models.

# Material and methods

We extracted clinical, pathologic and socioeconomic data of the patients in the Surveillance, Epidemiology, and End Results-18 (SEER-18) Program (www.seer.cancer.gov) SEER*Stat Database with Treatment Data, who were diagnosed of lung carcinoma in 2004. The follow up was last conducted in Nov. 2016 (the last in the 2019 data release). The inclusion criteria included the survival time > 1 month, aged 20+ years, first primary-cases only and having a known COD. The average follow-up time was 12.5 years.

The informed consent was not obtained for the SEER patients due to de-identified nature of the dataset. Owing to the use of publicly available, de-identified SEER cases, this study was exempt from an institutional review board approval. However, we have received the

approval for using the SEER-18 data under the condition of compliance with their preset terms (user ID lzhang). Moreover, all 50 states in the USA have laws requiring newly diagnosed cancers to be reported to a central registry. The state cancer registries in the SEER program would deposit their extracted, de-identified cancer data to the SEER database after meeting quality control standards (www.seer.cancer.gov). Thus, the SEER data collection was authorized by the US state laws, and supervised by respective state public-health officials and ethical review committees.

We simplified the SEER COD, which were extracted from death certificates, into 5 categories, namely cardiovascular system disease (CVS), infection, non-breast cancer, breast cancer and others cause (including alive). The alive patients were a small proportion of the population and were thus included in the other causes. We included 40 dichotomized variables in the analyses since dichotomized variables (i.e., one-hot encoding) produced slightly better accuracy as shown before[8] and would not require normalization.

We tuned and employed a random forest algorithm (Stata, version 15) to classify and predict specific causes of death in lung cancer patients. The lung cancer diagnosed during 2004 were included for the completeness in their follow-up and death causes. The patients were randomly divided into training and validation sets (1:1 match). We also compared the accuracies of the final random forest and multinomial regression models using chi-squared test. A p value $<0.05$ was considered statistically different.

## Results

Among the 44,735 case diagnosed in 2004, we identified and randomly selected 40,000 qualified lung cancers for the analyses, including 20,000 cases for either train or test set (**Table 1**). The causes of death were, in descending ranking order, were lung cancer (72.45%), other causes or alive (14.38%), non-lung cancer (6.87%), cardiovascular disease (5.35%), and infection (0.95%). The mean follow-up time was 114.8 months (standard deviation 49.1).

Table 1. Baseline characteristics of the included cases according to the 5-category Cause of death.

| | | **CVS,** n (%) | **Infection,** n (%) | **Lung cancer,** n (%) | **Non-lung cancer,** n (%) | **Other cause,** n (%) | **Total** | P* |
|---|---|---|---|---|---|---|---|---|
| **Case number** | | 2,261 | 407 | 30,597 | 2,896 | 6,096 | 42,257 | |
| **Age 65+ yr** | | | | | | | | <.001 |
| No | | 407 (18) | 124 (30) | 10330 (34) | 896 (31) | 2451 (40) | 14,208 | |
| Yes | | 1854 (82) | 283 (70) | 20267 (66) | 2000 (69) | 3645 (60) | 28,049 | |
| **Sex** | | | | | | | | <.001 |
| Female | | 957 (42) | 166 (41) | 13854 (45) | 1402 (48) | 3206 (53) | 19,585 | |
| Male | | 1304 (58) | 241 (59) | 16743 (55) | 1494 (52) | 2890 (47) | 22,672 | |
| **Grade 2tier** | | | | | | | | <.001 |
| Low | | 512 (43) | 91 (44) | 4543 (32) | 517 (39) | 2001 (53) | 7,664 | |
| High | | 689 (57) | 116 (56) | 9727 (68) | 814 (61) | 1769 (47) | 13,115 | |
| **T Category** | | | | | | | | <.001 |
| T1-2 | | 1074 (49) | 197 (49) | 7107 (24) | 903 (32) | 3522 (59) | 12,803 | |
| T3-4 | | 337 (15) | 59 (15) | 5607 (19) | 369 (13) | 748 (13) | 7,120 | |
| Unknown | | 793 (36) | 145 (36) | 17369 (58) | 1539 (55) | 1684 (28) | 21,530 | |
| **N category** | | | | | | | | <.001 |
| | 0 | 962 (43) | 175 (44) | 5485 (18) | 790 (28) | 3077 (51) | 10,489 | |
| | 1 | 120 (5) | 22 (6) | 1385 (5) | 108 (4) | 425 (7) | 2,060 | |
| | 2 | 315 (14) | 50 (13) | 5020 (17) | 322 (11) | 694 (12) | 6,401 | |
| | 3 | 37 (2) | 7 (2) | 1021 (3) | 71 (3) | 103 (2) | 1,239 | |
| Unknown | | 793 (36) | 145 (36) | 17369 (57) | 1539 (54) | 1684 (28) | 21,530 | |
| **M category** | | | | | | | | <.001 |
| | 0 | 1447 (65) | 260 (64) | 13226 (43) | 1312 (46) | 4334 (72) | 20,579 | |
| Unknown | | 793 (35) | 145 (36) | 17369 (57) | 1539 (54) | 1684 (28) | 21,530 | |
| **Summary stage 2000 (1998+)** | | | | | | | | <.001 |
| Blank(s) | | - | - | - | 16 (1) | - | 37 | |
| Distant | | 670 (30) | 121 (30) | 18190 (59) | 1490 (51) | 1221 (20) | 21,692 | |
| Localized | | 794 (35) | 141 (35) | 3686 (12) | 624 (22) | 2740 (45) | 7,985 | |
| Regional | | 667 (28) | 114 (28) | 7362 (24) | 639 (22) | 1905 (31) | 10,687 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Unknown/unstaged | 129 (6) | 23 (6) | 1352 (4) | 127 (4) | 225 (4) | 1,856 | |
| **Histology type** | | | | | | | <.001 |
| Adenocarcinoma | 544 (24) | 88 (22) | 8147 (27) | 837 (29) | 1718 (28) | 11,334 | |
| Small cell carcinoma | 143 (6) | 30 (7) | 4371 (14) | 264 (9) | 331 (5) | 5,139 | |
| Squamous cell carci.. | 524 (23) | 84 (21) | 5525 (18) | 456 (16) | 1092 (18) | 7,681 | |
| [OTHER] | 1050 (46) | 205 (50) | 12554 (41) | 1339 (46) | 2955 (48) | 18,103 | |
| **Laterality** | | | | | | | <.001 |
| Bilateral | 26 (1) | - | 573 (2) | 96 (3) | 50 (1) | 753 | |
| Left | 951 (42) | 168 (41) | 12529 (41) | 1175 (41) | 2482 (41) | 17,305 | |
| Right | 1284 (57) | 231 (57) | 17495 (57) | 1625 (56) | 3564 (58) | 24,199 | |
| **Diagnosis confirmation** | | | | | | | <.001 |
| Microscopic diagnosis | 1975 (87) | 346 (85) | 28339 (93) | 2628 (91) | 5611 (92) | 38,899 | |
| Radiologic and clinical diagnosis | 276 (12) | 61 (15) | 2090 (7) | 256 (9) | 468 (8) | 3,151 | |
| [OTHER] | - | 0 | 168 (1) | 12 (<1) | 17 (<1) | 207 | |
| **Radiotherapy** | | | | | | | <.001 |
| No | 1686 (75) | 307 (75) | 17869 (58) | 2039 (70) | 4790 (79) | 26,691 | |
| Yes | 575 (25) | 100 (25) | 12728 (42) | 857 (30) | 1306 (21) | 15,566 | |
| **Chemotherapy** | | | | | | | <.001 |
| No | 1669 (74) | 309 (76) | 15961 (52) | 1858 (64) | 4402 (72) | 24,199 | |
| Yes | 592 (26) | 98 (24) | 14636 (48) | 1038 (36) | 1694 (28) | 18,058 | |
| **Percent of high school education attainment, quartile§** | | | | | | | <.001 |
| Q1 | 575 (25) | 92 (23) | 7700 (25) | 762 (26) | 1676 (27) | 10,805 | |
| Q2 | 555 (25) | 115 (28) | 7873 (26) | 776 (27) | 1675 (27) | 10,994 | |
| Q3 | 592 (26) | 101 (25) | 7751 (25) | 738 (25) | 1394 (23) | 10,576 | |
| Q4 | 539 (24) | 99 (24) | 7273 (24) | 620 (21) | 1351 (22) | 9,882 | |
| **Percent of persons in poverty, quartile§** | | | | | | | <.001 |
| Q1 | 588 (26) | 92 (23) | 7544 (25) | 811 (28) | 1723 (28) | 10,758 | |
| Q2 | 532 (24) | 97 (24) | 7589 (25) | 728 (25) | 1512 (25) | 10,458 | |
| Q3 | 591 (26) | 102 (25) | 7921 (26) | 682 (24) | 1532 (25) | 10,828 | |
| Q4 | 550 (24) | 116 (29) | 7543 (25) | 675 (23) | 1329 (22) | 10,213 | |
| **Percent of foreign-born residents, quartile§** | | | | | | | <.001 |
| Q1 | 551 (24) | 117 (29) | 7958 (26) | 742 (26) | 1351 (22) | 10,719 | |
| Q2 | 573 (25) | 100 (25) | 7503 (25) | 725 (25) | 1592 (26) | 10,493 | |
| Q3 | 601 (27) | 104 (26) | 8311 (27) | 738 (25) | 1736 (28) | 11,490 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Q4 | 536 (24) | 86 (21) | 6825 (22) | 691 (24) | 1417 (23) | 9,555 |

Note: -, statistically suppressed due to fewer than 10 cases; *, chi-squared test; **§,** County attributes of Year 2000; Education attainment defined as percent of residents with less than high-school graduate in the county; Person in poverty defined as percent of residents with income below 200% of poverty in the county.

We found more 250 iterations and the 10 variables produced the best prediction, whose best accuracy was 69.8% (error-rate 30.2%, **Figure 1**).

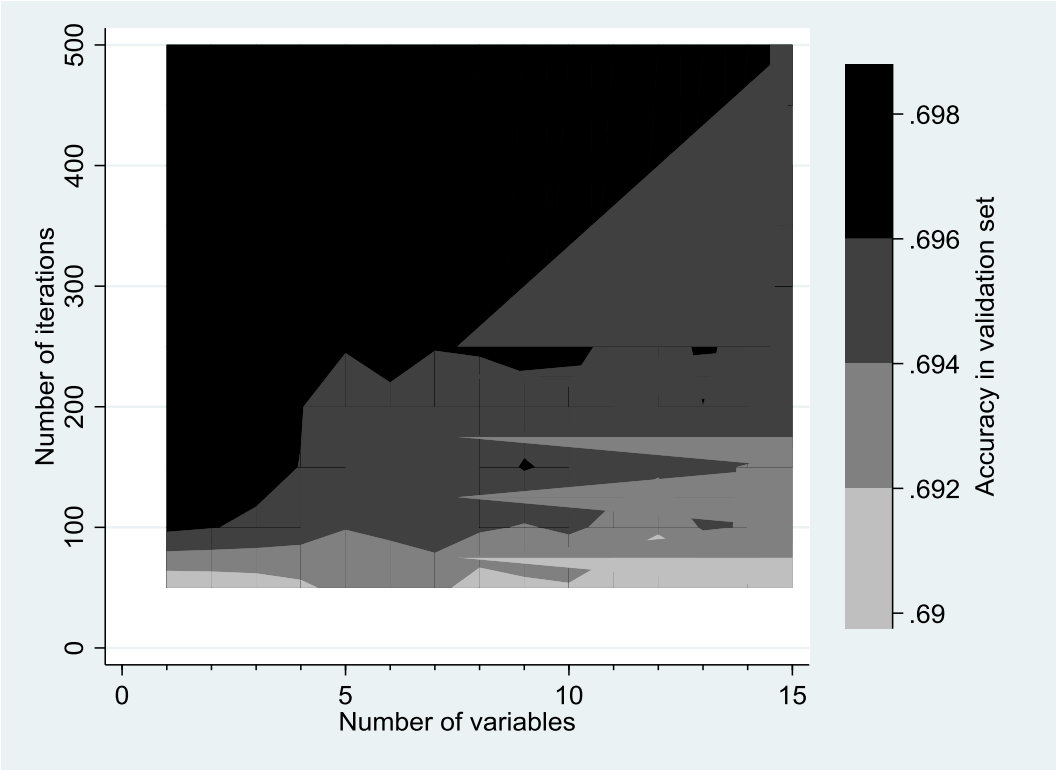


**Figure 1.** The error rates in the validation set were reduced as the number of iterations and variables increase. But more than 10 variables were linked to a higher error rate.

The final random forest model with 300 iterations and 10 variables reached an accuracy higher than that of multinomial regression model (69.8% vs 64.6%, P<0.001, **Table 2**). The top-10 most important factors in the random-forest model were sex, chemotherapy status, age (65+ vs <65 years), radiotherapy status, nodal status, T category, histology type and laterality (**Figure 2**).

Table 2. Confusion matrices of the random forest and multinomial regression models.

| Predicted classes | CVS | Infection | Lung cancer | Non-lung cancer | Other cause | Total |
|---|---|---|---|---|---|---|
| Random forest model | | | | | | |
| CVS | **3.68%** | 3.08% | 1.50% | 2.27% | 2.67% | 1.85% |
| Infection | 0.25% | **0.77%** | 0.23% | 0.10% | 0.37% | 0.25% |
| Lung cancer | 80.15% | 85.38% | **90.76%** | 85.66% | 69.54% | 86.72% |
| Non-lung cancer | 2.45% | 0.77% | 1.53% | **2.27%** | 2.67% | 1.79% |
| Other cause | 13.48% | 10.00% | 5.98% | 9.69% | **24.76%** | 9.39% |
| Multinomial regression models | | | | | | |
| Missing COD | 29.78% | 26.15% | 10.06% | 17.51% | 34.84% | 15.36% |
| Lung Cancer | 69.24% | 72.31% | **89.12%** | 81.40% | 58.67% | 82.96% |
| Multiple COD | 0.98% | 1.54% | 0.83% | 1.09% | 6.49% | 1.68% |

Note: COD, causes of death.

Interestingly, the pathological confirmation of the diagnosis appeared not very importance in the model. In the multinomial regression model, we also found that sex, chemotherapy status, age (65+ vs <65 years), radiotherapy status, nodal status, T category, histology type and laterality (**Table 3**), while other factors were not, despite being associated with the multicategory COD in univariable regression analysis (Chi-squared test, Table 1). This finding was consistent with that from RF model.

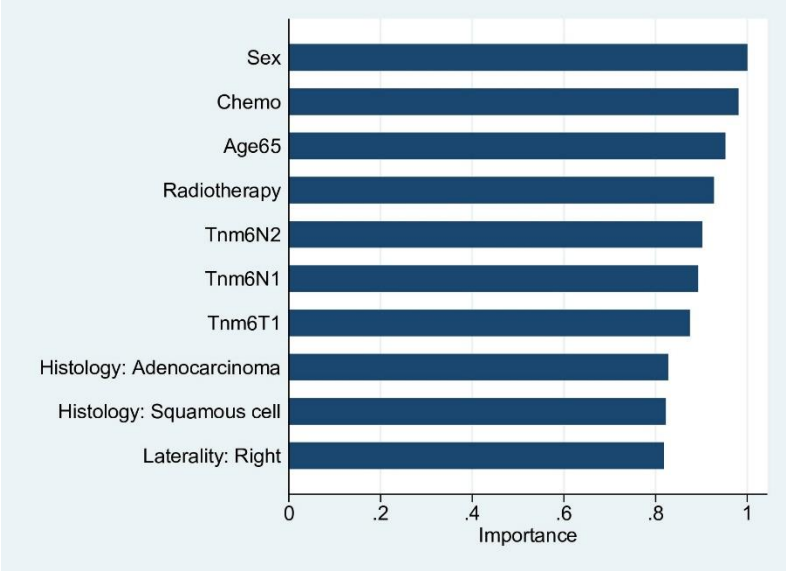

**Figure 2.** In the final random forest model of machine learning with 300 iteration, 10 variables and the highest accuracy (69.72%), the top-10 most important factors in the random-forest model were sex, chemotherapy status, age, radiotherapy status and nodal status. Confirmation by pathology was not in the selected 10 variables/factors.

Table 3. The factors associated with multicategory causes of death in lung cancer patients as shown in a multinomial regression model.

| | CVS | | | Infection | | | Non_lung_cancer | | | Other cause | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OR | 95% CI | P | OR | 95% CI | P | OR | 95% CI | P | OR | 95% CI | P |
| **Male (vs female)** | 1.19 | (1.04 - 1.36) | 0.009 | 1.21 | (0.90 - 1.63) | 0.197 | 0.95 | (0.85 - 1.07) | 0.397 | 0.79 | (0.73 - 0.86) | <.001 |
| **Radiotherapy (vs none)** | 0.70 | (0.60 - 0.82) | <.001 | 0.47 | (0.32 - 0.71) | <.001 | 0.76 | (0.67 - 0.86) | <.001 | 0.53 | (0.47 - 0.59) | <.001 |
| **Chemotherapy (vs none)** | 0.61 | (0.51 - 0.71) | <.001 | 0.59 | (0.41 - 0.87) | 0.007 | 0.76 | (0.66 - 0.87) | <.001 | 0.60 | (0.54 - 0.67) | <.001 |
| **Age (65+ vs <65 yr)** | 1.83 | (1.54 - 2.17) | <.001 | 0.73 | (0.53 - 1.00) | 0.047 | 1.00 | (0.88 - 1.13) | 0.964 | 0.56 | (0.51 - 0.62) | <.001 |
| **Histology** | | | | | | | | | | | | |
| **Adenocarcinoma** | 0.73 | (0.61 - 0.86) | <.001 | 0.64 | (0.43 - 0.95) | 0.025 | 0.99 | (0.86 - 1.14) | 0.903 | 0.73 | (0.65 - 0.81) | <.001 |
| **Small cell carcinoma** | 0.63 | (0.48 - 0.84) | 0.001 | 0.84 | (0.47 - 1.49) | 0.554 | 0.67 | (0.54 - 0.83) | <.001 | 0.49 | (0.40 - 0.59) | <.001 |
| **Squamous cell carcinoma** | 1.12 | (0.95 - 1.34) | 0.183 | 0.95 | (0.64 - 1.41) | 0.797 | 0.82 | (0.70 - 0.98) | 0.026 | 0.75 | (0.67 - 0.85) | <.001 |
| **Other Histology** | - | | - | - | | - | - | | - | - | | - |
| **N category** | | | | | | | | | | | | |
| **N1 vs N0** | 3.18 | (1.65 - 6.12) | 0.001 | 7.51 | (1.38 - 40.90) | 0.02 | 1.43 | (0.87 - 2.34) | 0.158 | 2.08 | (1.40 - 3.09) | <.001 |
| **N2 vs N0** | 2.42 | (1.19 - 4.90) | 0.014 | 2.86 | (0.42 - 19.51) | 0.284 | 1.42 | (0.81 - 2.49) | 0.215 | 1.79 | (1.16 - 2.77) | 0.008 |
| **N3 vs N0** | 1.91 | (1.04 - 3.50) | 0.036 | 1.74 | (0.40 - 7.53) | 0.456 | 0.87 | (0.56 - 1.33) | 0.516 | 1.36 | (0.96 - 1.91) | 0.079 |
| **Percent of foreign-born residents, quartile** | | | | | | | | | | | | |
| **Q1 vs Q4** | 0.89 | (0.71 - 1.11) | 0.298 | 0.65 | (0.39 - 1.08) | 0.093 | 1.01 | (0.83 - 1.22) | 0.955 | 0.79 | (0.67 - 0.92) | 0.002 |
| **Q2 vs Q4** | 0.96 | (0.77 - 1.20) | 0.724 | 0.81 | (0.50 - 1.34) | 0.415 | 0.95 | (0.78 - 1.15) | 0.613 | 0.93 | (0.80 - 1.08) | 0.349 |
| **Q3 vs Q4** | 0.81 | (0.67 - 0.98) | 0.035 | 0.78 | (0.51 - 1.19) | 0.242 | 0.84 | (0.71 - 0.99) | 0.037 | 0.91 | (0.80 - 1.03) | 0.14 |
| **Laterality** | | | | | | | | | | | | |
| **Bilateral (vs Right side)** | 0.66 | (0.32 - 1.36) | 0.261 | 2.49 | (1.06 - 5.88) | 0.037 | 1.62 | (1.14 - 2.31) | 0.007 | 0.83 | (0.53 - 1.28) | 0.397 |
| **Left side (vs Right side)** | 1.06 | (0.93 - 1.21) | 0.404 | 0.84 | (0.63 - 1.14) | 0.272 | 0.96 | (0.86 - 1.08) | 0.504 | 0.99 | (0.91 - 1.09) | 0.884 |

Note: OR, odds ratio; CI, confidence interval.

## Discussion

In this population-based study of 40,000 lung cancer patients with more than 12 years of follow-up, we optimized a RF model of machine learning to predict the specific cause of death in lung cancer patients. The RF model also appears more accurate than multinomial regression model (69.8% vs 64.6%, P<0.001). We also identified the factors

linked to the 5-category COD in lung cancer patients, which was not reported by previous works [4, 7].

This study is limited by using the cases of a single diagnosis year, while we feel the large number of cases justified for our findings. Future studies may include the cases of recent diagnosis years so that the findings will be more related to the recent cases. Moreover, the dataset appeared unbalanced in the outcome. Such a nature of the dataset is the real-world evidence, but posed a challenge to modelling. This in part explained why the multinomial regression model had many cases missing an assigned COD. Finally, the findings were not validated by an external dataset. Future works should be carried out using other dataset to validate our findings.

In summary, we here identified the factors that were independently associated with 5-category long-term COD in lung cancer patients in the USA. We also tuned the RF model and showed that it was significantly more accurate in predicting 5-category long-term COD.

## Data availability

The SEER data were available upon request to the SEER website (www.seer.cancer.gov). All other data are available upon request.

## References

1. Siegel RL, Miller KDJemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020; 70: 7-30.
2. Parikh N, Shuck RL, Gagea M, Shen LDonehower LA. Enhanced inflammation and attenuated tumor suppressor pathways are associated with oncogene-induced lung tumors in aged mice. Aging Cell. 2018; 17:
3. Shang G, Jin Y, Zheng Q, Shen X, Yang M, Li YZhang L. Histology and oncogenic driver alterations of lung adenocarcinoma in Chinese. Am J Cancer Res. 2019; 9: 1212-1223.
4. Yin J, Zhao M, Lu T, Huang Y, Sui Q, Xi J, Lin Z, Xu S, Wang QZhan C. Non-lung cancer specific mortality after lobectomy or sublobectomy in patients with stage IA non-small cell lung cancer $\leqslant$2 cm: A propensity score analysis. J Surg Oncol. 2019; 120: 1486-1496.
5. Sturgeon KM, Deng L, Bluethmann SM, Zhou S, Trifiletti DM, Jiang C, Kelly SPZaorsky NG. A population-based study of cardiovascular disease mortality risk in US cancer patients. Eur Heart J. 2019; 40: 3889-3897.
6. Gad MM, Saad AM, Al-Husseini MJ, Abushouk AI, Salahia S, Rehman KA, Riaz HAhmed HM. Temporal trends, ethnic determinants, and short-term and long-term risk of cardiac death in cancer patients: a cohort study. Cardiovasc Pathol. 2019; 43: 107147.
7. Zaorsky NG, Churilla TM, Egleston BL, Fisher SG, Ridge JA, Horwitz EMMeyer JE. Causes of death among cancer patients. Ann Oncol. 2017; 28: 400-407.
8. Wang J, Deng F, Zeng F, Shanahan AJ, Li WVZhang L. Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. Am J Cancer Res. 2020; 10: 1344-1355.