Materials and methods

A Fisher Test, Logistic Regression, Cox Regression analysis, Kaplan-Meier, Logrank Test

B Radiologic, clinical, and pathologic profiles of surgically confirmed SCCs from 398 patients

Survival Analysis

1 May 2020

Evaluation of Clinicopathological and Radiologic factors on lung squamous cell carcinoma (SCC)

Khwansiri NINPAN

Lisa KOPPE

Princy PAPPACHAN

Objective:

Evaluate clinicopathological and radiologic factors for the prognosis of lung squamous cell carcinoma.

Data description:

This dataset contains 398 records from patients who had curative surgery for lung squamous cell carcinoma (SCC) at Samsung Medical Center, Korea between January 2003 to December 2012*.
The records of the 21 features can be separated into 3 groups: Clinical, Histopathologic and Radiological features. Demographic details are listed in Table 1.

Table 1. Patient characteristics

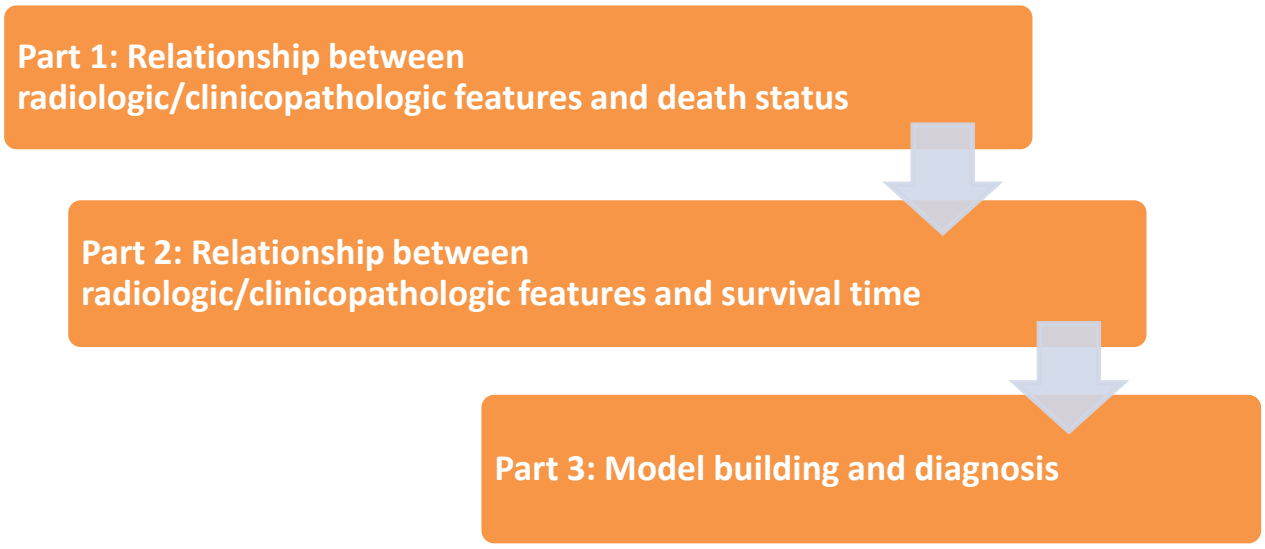
	Feature Name	Variable Name in code	Type	Variable Distribution
Clinical features	Age at diagnosis (years)**	AG (LT60, OV60)		66 [60,71]
	Sex	M	Male	384 (96.5%)
		F	Female	14 (3.5%)
	Death status	0	Alive	226 (56.8%)
		1	Death	172 (43.2%)
	Overall Survival Time (months)**	OS		53 (24, 83)
	Smoking state of the patient	1	Current smoker	116 (29.1%)
		2	Ex-smoker	202 (50.7%)
		3	Non-smoker	23 (5.8%)
		0	Unknown	57 (14.3%)
	Pulmonary function test (% FEV1/FVC)**	FEV1_FVC		54 [47,60]
	Serum carcinoembryonic antigen (ng/ml)**	CEA		2.2 [1.37-3.322]
	Metastasis	0	Absent	236 (59.3%)
		1	Present	160 (40.2%)
Histopathologic features	Lymph node ratio***	LN_ratio		0.045 ± 0.14
	Criteria for size of original tumor in Central Lung SCC	cT	cT1	73 (18%)
			cT2	252 (63.3%)
			cT3	58 (14.5%)
			cT4	15 (3.7%)
	Criteria for size of original tumor in Peripheral Lung SCC	pT	pT1	72 (18%)
			pT2	240 (60.3%)
			pT3	78 (19.6%)
			pT4	8 (0.2%)
	Criteria for nearby lymph nodes involved in Central Lung SCC	cN	cN0	284(71.3%)
			cN1	68 (17%)
			cN2	39 (9.7%)
			cN3	7 (1.8%)
	Criteria for nearby lymph nodes involved in Peripheral Lung SCC	pN	pN0	246 (61.8%)
			pN1	95 (23.9%)
			pN2	53 (13.3%)
			pN3	4 (1%)
	Criteria for distant metastasis involved in Central Lung SCC	cM	cM0	396 (99.5%)
			cM1b	2 (0.5%)
	Criteria for distant metastasis involved in Peripheral Lung SCC	pM	pM0	394 (98.9%)
			pM1b	4 (1%)

Table 1. Patient characteristics (cont.)

	Feature Name	Variable Name in Code	Type	Variable Distribution
Radiological features	Cancer location	0	Peripheral	143 (35.9%)
		1	Central	255 (64.1%)
	Present of obstructive pneumonitis/atelectasis	0	Absent	206 (51.8%)
		1	Present	192 (48.2%)
	Necrosis	0	No Remark	257 (64.6%)
		1	Necrosis	116 (29.1%)
		2	Cavitation	25 (6.3%)
	Underlying lung disease	0	No Remark	292 (73.4%)
		1	Emphysema	85 (21.3%)
		2	Interstitial lung abnormality (ILA)	21 (5.3%)
	Lung effusion	0	Absent	389 (97.8%)
		1	Present	9 (2.2%)

**Ref: Integrated evaluation of clinical, pathological and radiological prognostic factors in squamous cell carcinoma of the lung, PLoSOne. 2019 Oct 4;14(10):e0223298. doi: 10.1371/journal.pone.0223298. eCollection 2019.*
*** Data are median values [1st quartile, 3rd quartile]*
**** Data are mean values ± standard deviation*

Project design:



Part 1: Relationship between radiologic/clinicopathologic features and death status

Objective: Define the features which affect the death prognosis of lung squamous cell carcinoma.

Methods used:

1. Fisher's exact test: determine the relationships between two categorical variables.
2. Logistic regression: determine the relationships between categorical and continuous variables.

Results and discussion:

Table 2: Relationships between radiologic/clinicopathologic features and death (univariate analysis)

Variables	Comparison group(s)	Reference group	P-value	Odds ratio
Age	**Age over than 60 years old**	Age less than 60 years old	**0.002**	**2.214**
Sex	Male	Female	0.5974	0.754
Smoking state	Current smoker	Non-smoker	0.076	2.539
	Ex-smoker		0.3673	1.557
	Unknown		0.2318	1.764
Pulmonary function test (% FEV1/FVC)	numerical value		0.2112	1.011
Serum carcinoembryonic antigen (ng/ml)	numerical value		0.5811	0.997
Metastasis	**Present**	Absent	**8.52e-10**	**3.656**
Lymph node ratio	numerical value		**0.023**	**3.975**
Criteria for size of original tumor in Central Lung SCC	**cT2**	cT1	**0.048**	**1.769**
	cT3		**9e-05**	**4.363**
	cT4		**0.0254**	**3.714**
Criteria for nearby lymph nodes involved in Central Lung SCC	cN1	cN0	0.948	0.982
	cN2		**0.001**	**3.355**
	cN3		0.374	1.988
Criteria for distant metastasis involved in Central Lung SCC	cM1b	cM0	0.847	1.316
Criteria for size of original tumor in Peripheral Lung SCC	pT2	pT1	0.125	1.548
	pT3		**0.005**	**2.617**
	pT4		0.101	3.551
Criteria for nearby lymph nodes involved in Peripheral Lung SCC	pN1	pN0	0.365	1.249
	pN2		**0.0003**	**3.20**
	pN3		0.170	4.94
Criteria for distant metastasis involved in Peripheral Lung SCC	pM1b	pM0	0.983	
Cancer location	Central	Peripheral	0.1914	1.316
Present of obstructive pneumonitis/atelectasis	Present	Absent	0.3622	0.816
Necrosis	Necrosis	No Remark	0.310	1.258
	Cavitation		**0.025**	**2.658**
Underlying lung disease	Emphysema	No Remark	0.3893	1.669
	Interstitial lung abnormality (ILA)		**0.0003**	**9.784**
Lung effusion	Present	Absent	1	1.052

Based on the univariate analysis with Fisher's exact test and logistic regression, we found features that show statistically significant relationship with the risk of death as follow: age, metastasis, lymph node ratio, tumor size, nearby lymph node involvement, the present of cavitation in lung and underlying lung disease with interstitial lung abnormality (ILA).

To define the relationship between age and risk of death, we divide the patients into 2 age groups, less than and over 60 years old. Fisher's test has p-value **0.002** which means the difference in the death between these 2 age groups is statistically significant. Odds ratio **2.214** suggests that the patients who are older than 60 years old have twice higher risk to death than patients who are younger. *(see code in supporting information)*

The p-value of Fisher's test between the presence of metastasis after surgery and death is way lower than the significant level of **5%** (p-value=**8.52e-10**) the variables are closely dependent. Data is sufficient to prove that the observation of metastasis after surgery increases the risk leading to death. The odds ratio (**3.656**) suggests that patients with observed metastasis after surgery have approximately 4 times higher risk of death than patients without any metastasis observed.

Lymph node ratio shows positive coefficient in logistic regression analysis which means that each unit increase of lymph node ratio has a positive association with approximately 4 times higher risk of death (p-value=**0.023**).

The increasing size of the tumor in lung cancer has positive association with the risk of death. For central lung cancer, tumor size T2, T3 and T4 have approximately **77%**, **4.4** times and **3.7** times more chances of death as compared to T1 with p-value **0.048**, **9e-05** and **0.0254** respectively. While in peripheral lung cancer, only tumor size T3 shows statistically significant higher risk of death (approximately **3** times higher) with p-value **0.005**.

Nearby lymph node involvement is also associate with the risk of death. In both central and peripheral lung cancer, patients who have criteria N2 have approximately **3** times higher risk of death than that of criteria N0. This difference is statistically significant with p-value **0.001** and **0.0003** in central and peripheral lung SCC, respectively.

The present of lung cavitation and interstitial lung abnormality (ILA) have statistically significant involvement in risk of death. They have approximately **3** times (p-value=**0.025**) and **10** times (p-value=**0.0003**) higher risk of death than those without cavitation and ILA condition, respectively. *(see code in supporting information)*

The features that show no statistically significant relationship with death are sex (p-value=**0.5974**), pulmonary function test (p-value=**0.2112**), serum carcinoembryonic antigen (p-value=**0.5811**), current smoker (p-value=**0.0759**), ex-smoker (p-value=**0.3673**), distant metastasis (p-value=**0.847** in central lung cancer and **0.983** in peripheral lung cancer), cancer location (p-value=**0.2**), the present of obstructive pneumonitis/atelectasis (p-value=**0.3622**) and lung effusion (p-value=**1**).

Part 2: Relationship between radiologic/clinicopathologic features and survival time (univariate analysis)

Objective: Define the features which affect the prognosis of lung squamous cell carcinoma survival time.

Methods used:

1. Cox Proportional Hazards Regression Model: explore the relationships between the survival time and explanatory variables.
2. Kaplan-Meier estimator: estimate the survival function and visualize probability of an event at a certain time interval.
3. Logrank test: compare the survival between groups of populations.

Results and discussion:
Relationship between each features and death status is showed in Table 3.

Part2. A. Cox Proportional Hazards Regression Model fit

The Cox regression results can be interpreted as follow:

1. Statistical significance:
The Wald statistic value z corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The Wald statistic evaluates whether the beta (β) coefficient of a given variable is statistically significantly different from 0.

2. The regression coefficient:
The second feature to note in the Cox model results is the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group.

3. Hazard ratios (HR):
The exponentiated coefficient (exp(coef)) gives the effect size of covariates.

4. Confidence intervals of the hazard ratios:
The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)).

5. Global statistical significance of the model:
Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score Logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

Thanks to these results, we can decide to reject or keep covariates for the upcoming study based on their statistical significance.

We here split the analysis into two groups of features:
- Radiologic features
- Clinicopathologic features

Part2. A. 1. Relationship between radiologic features and survival time

Features to be analyzed in this part:

1. Location vs OS: tumor location
2. Obst_pn_or_plugging vs OS: patients with or without obstructive pneumonitis/atelectasis
3. Necrosis vs OS: patients with or without cell necrosis or cavitation observed
4. Underlying_lung vs OS: the type of underlying lung disease observed
5. Effusion vs OS: patient with or without observed lung effusion

The p-values of Location (0.09), Obst_pn_or_plugging (0.1), Necrosis (0.05) and effusion (0.9) are too high to consider these features as statistically significant. We then reject the use of these four covariates for the upcoming study. *(see code in supporting information)*

However, with a very low p-value of **3.93e-10** and a positive coefficient of **1.61**, we can assume that there is a highly significant positive relationship between survival time (OS) and Interstitial lung abnormalities (ILA). With a value of **1.61**, the sign of the regression coefficient for ILA is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for ILA=**1.61** indicates that patients with ILA have higher risk of death (lower survival rates) than patients with emphysema. The exponentiated coefficient (exp(coef)=exp(1.61)=**5.00**) gives the effect size of covariates. Here, a patient with ILA increases the hazard by a factor of **5.00**, or **400%**. Having ILA is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**3.02**, upper 95% bound=**8.28** for ILA. *(see code in supporting information)*

To conclude, Underlying_lung is considered statistically significant.

Part2. A. 2. Relationship between clinicopathologic features and survival time

Features to be analyzed in this part:

1. Sex vs OS: gender of the patient
2. AG vs OS: age of the patient at diagnosis
3. LN_ratio vs OS: lymph node ratio
4. cT vs OS: central lung SCC size of tumor
5. cN vs OS: when nearby lymph nodes are involved
6. cM vs OS: when distant metastasis are involved
7. pT vs OS: peripheral lung SCC size of tumor
8. pN vs OS: when nearby lymph nodes are involved

9. pM vs OS: when distant metastasis are involved
10. Smoking_state vs OS: current smoking state of the patient
11. FEV1_FVC vs OS: pulmonary function test
12. CEA vs OS: serum carcinoembryonic antigen level
13. OP vs OS: the type of operation undergone by the patient
14. meta vs OS: patient with and without observed metastasis after surgery

The p-values of Sex (0.5), cM (0.7), Smoking_state (0.2), FEV1_FVC (0.1) and CEA (0.6) are too high to consider these features as statistically significant. We then reject the use of these five covariates for the upcoming study. *(see code in supporting information)*

Also, the repartition of the population in the groups is too unbalanced for the LN_ratio and the OP covariates so these features will also be rejected for the next steps of the study. *(see code in supporting information)*

However, with a low p-value of **0.001** and a positive coefficient of **0.70**, we can assume that there is a statistically significant positive relationship between survival time (OS) and age over 60. With a value of **0.70**, the sign of the regression coefficient for age over 60 is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for age over 60=**0.70** indicates that patients aged over 60 have higher risk of death (lower survival rates) than younger patients. The exponentiated coefficient (exp(coef)=exp(0.70)=**2.01**) gives the effect size of covariates. Here, a patient older than 60 increases the hazard by a factor of **2.01**, or **101%**. Being older than 60 is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.33**, upper 95% bound=**3.06** for age over 60. *(see code in supporting information)*

With a low p-value of **2.99e-05** and a positive coefficient of **1.14**, we can assume that there is a statistically significant positive relationship between survival time (OS) and cT3. With a value of **1.14**, the sign of the regression coefficient for cT3 is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for cT3=**1.14** indicates that patients with cT3 have higher risk of death (lower survival rates) than patients with cT1. The exponentiated coefficient (exp(coef)=exp(1.14)=**3.14**) gives the effect size of covariates. Here, a patient with cT3 increases the hazard by a factor of **3.14**, or **214%**. Having cT3 is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.84**, upper 95% bound=**5.38** for cT3. *(see code in supporting information)*

With a low p-value of **0.0003** and a positive coefficient of **0.77**, we can assume that there is a statistically significant positive relationship between survival time (OS) and cN2. With a value of **0.77**, the sign of the regression coefficient for cN2 is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for cN2=**0.77** indicates that patients with cN2 have higher risk of death (lower survival rates) than patients with cN0. The exponentiated coefficient (exp(coef)=exp(0.77)=**2.16**) gives the effect size of covariates. Here, a patient with cN2 increases the hazard by a factor of **2.16**, or **116%**. Having cN2 is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.42**, upper 95% bound=**3.29** for cN2. *(see code in supporting information)*

With a low p-value of **0.007** and a positive coefficient of **0.69**, we can assume that there is a statistically significant positive relationship between survival time (OS) and pT3. With a value of **0.69**, the sign of the regression coefficient for pT3 is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for pT3=**0.69** indicates that patients with pT3 have higher risk of death (lower survival rates) than patients with pT1. The exponentiated coefficient (exp(coef)=exp(0.69)=**1.99**) gives the effect size of covariates. Here, a patient with pT3 increases the hazard by a factor of **1.99**, or **99%**. Having pT3 is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.20**, upper 95% bound=**3.30** for pT3. *(see code in supporting information)*

With a low p-value of **1.51e-05** and a positive coefficient of **0.86**, we can assume that there is a statistically significant positive relationship between survival time (OS) and pN2. With a value of **0.86**, the sign of the regression coefficient for pN2 is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for pN2=**0.86** indicates that patients with pN2 have higher risk of death (lower survival rates) than patients with pN0. The exponentiated coefficient (exp(coef)=exp(0.86)=**2.37**) gives the effect size of covariates. Here, a patient with pN2 increases the hazard by a factor of **2.37**, or **137%**. Having pN2 is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.60**, upper 95% bound=**3.50** for pN2. *(see code in supporting information)*

With a low p-value of **0.0005** and a positive coefficient of **1.77**, we can assume that there is a statistically significant positive relationship between survival time (OS) and pM1b. With a value of **1.77**, the sign of the regression coefficient for pM1b is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for pM1b=**1.77** indicates that patients with pM1b have higher risk of death (lower survival rates) than patients with pM0. The exponentiated coefficient (exp(coef)=exp(1.77)=**5.85**) gives the effect size of covariates. Here, a patient with pM1b increases the hazard by a factor of **5.85**, or **485%**. Having pM1b is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**2.16**, upper 95% bound=**15.84** for pM1b. *(see code in supporting information)*

With a low p-value of **5.22e-08** and a positive coefficient of **0.85**, we can assume that there is a statistically significant positive relationship between survival time (OS) and the observation of metastasis. With a value of **0.85**, the sign of the regression coefficient for observed metastasis is positive which means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The beta coefficient for observed metastasis=**0.85** indicates that patients with observed metastasis have higher risk of death (lower survival rates) than patients without metastasis. The exponentiated coefficient (exp(coef)=exp(0.85)=**2.33**) gives the effect size of covariates. Here, a patient with observed metastasis increases the hazard by a factor of **2.33**, or **133%**. Observed metastasis is indeed associated with bad prognostic. Moreover, the summary output also gives upper and lower 95% confidence intervals (CI) for the hazard ratio (exp(coef)) as follows: lower 95% bound=**1.72**, upper 95% bound=**3.16** for observed metastasis. *(see code in supporting information)*

To conclude, Underlying_lung , AG, cT, cN, pT, pN, pM, and meta are considered statistically significant.

Table 3. Relationship between radiologic/clinicopathologic features and survival time (univariate analysis)

Variables	Comparison group(s)	Reference group	P-value	HR [95% CI]
Location	Peripheral	Central	0.0891	1.304 [0.9602-1.771]
Obstructive pneumonitis/atelectasis	Present	Absent	0.099	0.7765 [0.575-1.049]
Necrosis	Necrosis	NoRemark	0.110	1.310 [0.9402-1.826]
	Cavitation		0.033	1.774 [1.0473-3.006]
Underlying lung disease	Emphysema	NoRemark	0.0255	1.494 [1.051-2.125]
	ILA		**3.93e-10**	**5.001 [3.020-8.280]**
Effusion	Present	Absent	0.917	0.9486 [0.3518-2.557]
Sex	M	F	0.481	0.7619 [0.3574-1.624]
Age	**OV60**	LT60	**0.00103**	**2.015 [1.326-3.062]**
Lymph node ratio	numerical value		0.00956	2.329 [1.229-4.414]
cT	cT2	cT1	0.1236	1.445 [0.9044-2.310]
	cT3		**2.99e-05**	**3.144 [1.8361-5.384]**
	cT4		0.0642	2.092 [0.9574-4.573]
cN	cN1	cN0	0.634426	1.107 [0.7276-1.685]
	cN2		**0.000321**	**2.162 [1.4205-3.291]**
	cN3		0.218450	1.872 [0.6896-5.085]
cM	cM1b	cM0	0.667	1.54 [0.2155-11]
pT	pT2	pT1	0.20657	1.339 [0.8512-2.106]
	pT3		**0.00787**	**1.988 [1.1976-3.299]**
	pT4		0.20356	1.873 [0.7118-4.930]
pN	pN1	pN0	0.1864	1.281 [0.887-1.851]
	pN2		**1.51e-05**	**2.367 [1.602-3.498]**
	pN3		0.0493	3.181 [1.004-10.084]
pM	**pM1b**	pM0	**0.000504**	**5.854 [2.163-15.84]**
Smoking state	Current smoker	Non-smoker	0.103	1.984 [0.8708-4.520]
	Ex-smoker		0.345	1.466 [0.6625-3.243]
	Unknown		0.122	1.838 [0.8507-3.969]
FEV1FVC	numerical value		0.131	1.01 [0.9969-1.024]
CEA	numerical value		0.604	0.9979
Operation type	Lobectomy	No op	0.0925	0.3007 [0.07415-1.219]
	Pneumonectomy		0.2076	0.3964 [0.09400-1.672]
	Segmentectomy		0.2143	0.2183 [0.01976-2.411]
	LN biopsy		0.6951	1.3522 [0.29915-6.112]
Metastasis	**Present**	Absent	**5.22e-08**	**2.332 [1.719-3.163]**

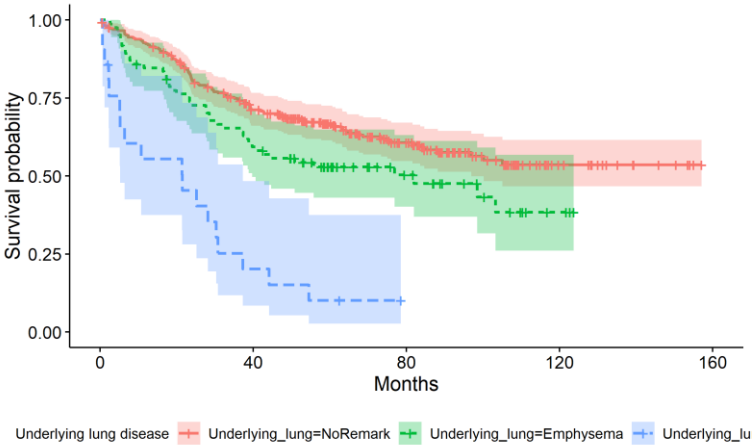
Part2. B. Kaplan-Meier estimator and Logrank test with features which have a significant relationship with survival time

List of features which have a significant relationship with survival time (OS):

1. Comparison between 2 groups of patients based on underlying lung disease
2. Comparison between 2 groups of patients based on their age
3. Comparison between 2 groups of patients based on central lung SCC size of tumor
4. Comparison between 2 groups of patients based on cN when nearby lymph nodes are involved
5. Comparison between 2 groups of patients based on peripheral lung SCC size of tumor
6. Comparison between 2 groups of patients based on pN when nearby lymph nodes are involved
7. Comparison between 2 groups of patients based on pM when distant metastasis are involved
8. Comparison between 2 groups of patients based on the observation of metastasis after surgery

The Logrank test result doesn't show a significant difference in survival between groups of patients according to the peripheral lung SCC size of tumor (p-value=0.04).
We then reject the use of the pT covariate. *(see code in supporting information)*

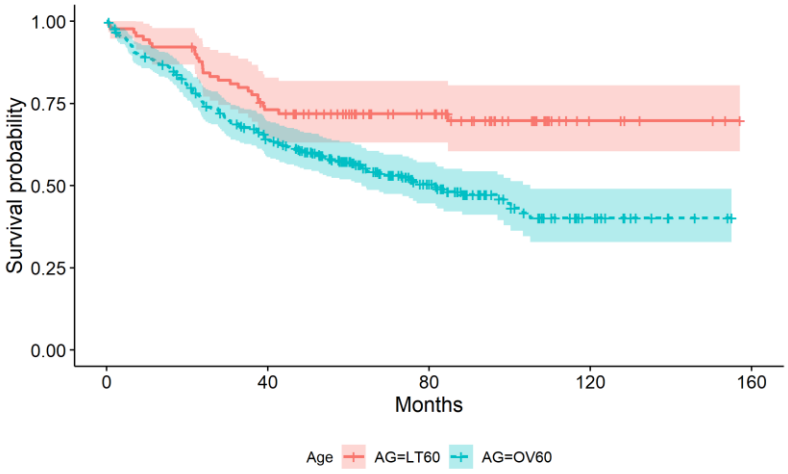
1. Comparison between 2 groups of patients based on underlying lung disease



Among a total of 292 patients without underlying lung disease, 111 of them died.
Among a total of 85 patients with emphysema, 43 of them died with 40% chance of death at 82 months.
Among a total of 21 patients with ILA, 18 of them died with 5% chance of death at 21.5 months.
Logrank test result shows a significant difference in survival between these groups of patients (p=4e-11).

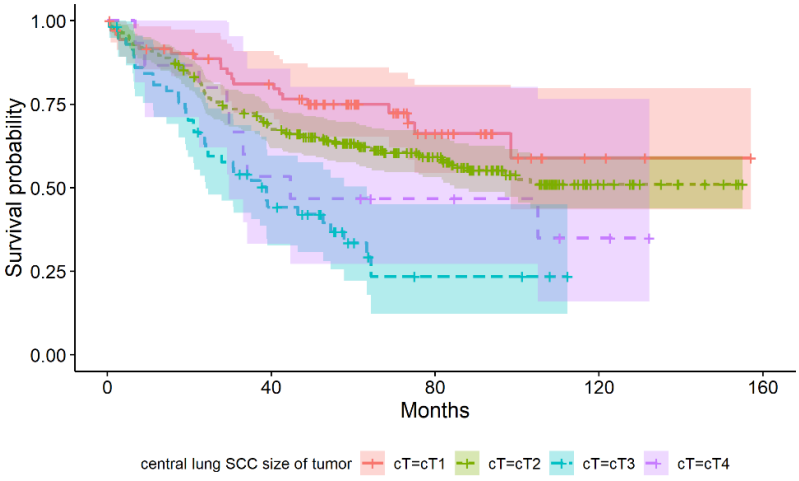
(see code in supporting information)

2. Comparison between 2 groups of patients based on their age



Among a total of 90 patients below the age of 60, 26 of them died.
Among a total of 308 patients above the age of 60, 146 of them died with 63% chance of death at 81.6 months.
Logrank test result shows a significant difference in survival between these groups of patients (p=8e-04).
(see code in supporting information)

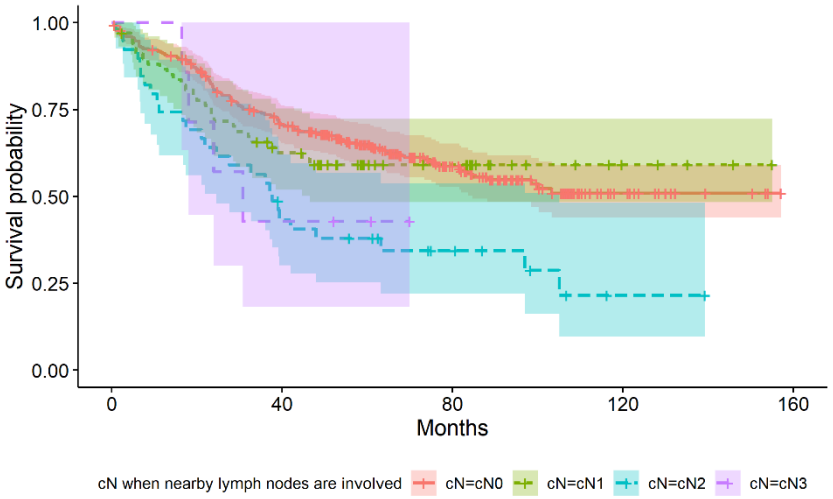
3. Comparison between 2 groups of patients based on central lung SCC size of tumor



Among a total of 73 patients with central lung SCC size of tumor T1, 21 of them died.
Among a total of 252 patients with central lung SCC size of tumor T2, 105 of them died.
Among a total of 58 patients with central lung SCC size of tumor T3, 37 of them died with 24% chance of death at 38.8 months.
Among a total of 15 patients with central lung SCC size of tumor T4, 9 of them died with 30% chance of death at 44.7 months.

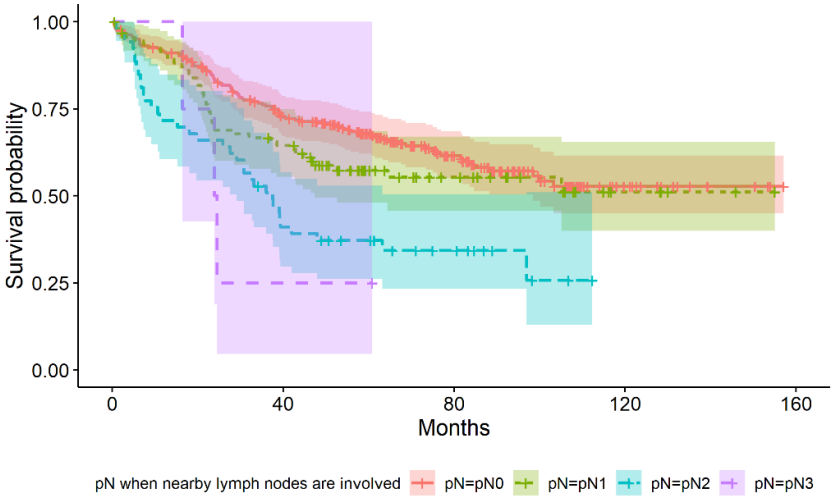
Logrank test result shows a significant difference in survival between these groups of patients (p=3e-05). *(see code in supporting information)*

4. Comparison between 2 groups of patients based on cN when nearby lymph nodes are involved



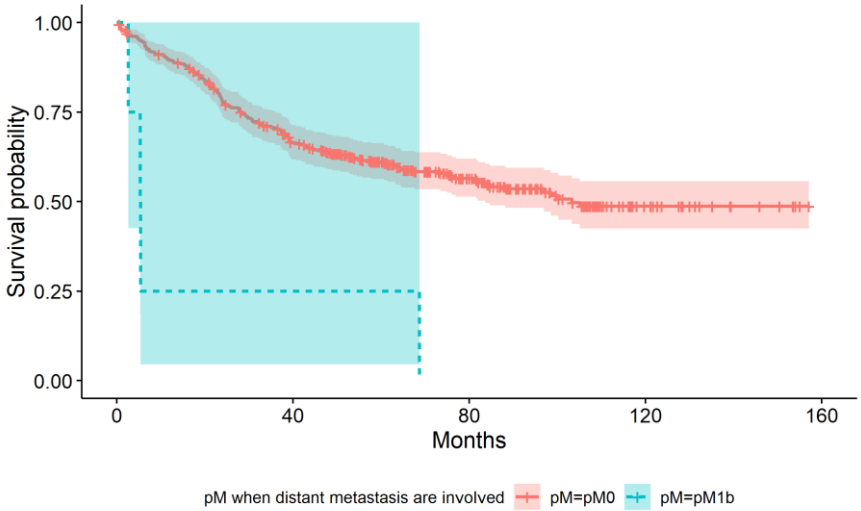
Among a total of 284 patients with N0 central nearby lymph nodes involved, 114 of them died.
Among a total of 68 patients with N1 central nearby lymph nodes involved, 27 of them died.
Among a total of 39 patients with N2 central nearby lymph nodes involved, 27 of them died with 25% chance of death at 37.6 months.
Among a total of 7 patients with N3 central nearby lymph nodes involved, 4 of them died with 18% chance of death at 30.8 months.
Logrank test result shows a significant difference in survival between these groups of patients (p=0.002). *(see code in supporting information)*

6. Comparison between 2 groups of patients based on pN when nearby lymph nodes are involved



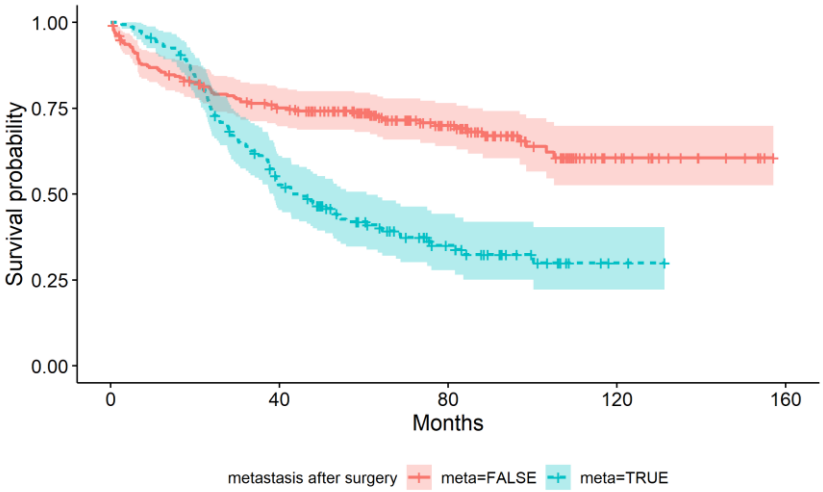
Among a total of 246 patients with N0 peripheral nearby lymph nodes involved, 93 of them died.
Among a total of 95 patients with N1 peripheral nearby lymph nodes involved, 41 of them died.
Among a total of 53 patients with N2 peripheral nearby lymph nodes involved, 35 of them died with 30% chance of death at 37.5 months.
Among a total of 4 patients with N3 peripheral nearby lymph nodes involved, 3 of them died with 16% chance of death at 24.2 months.
Logrank test result shows a significant difference in survival between these groups of patients (p=5e-05). *(see code in supporting information)*

7. Comparison between 2 groups of patients based on pM when distant metastasis are involved



Among a total of 394 patients with M0 peripheral distant metastasis involved, 168 of them died with 84% chance of death at 103.3 months.
Among a total of 4 patients with M1b peripheral distant metastasis involved, all of them died with 3% chance of death at 5.35 months.
Logrank test result shows a significant difference in survival between these groups of patients (p=8e-05).
(see code in supporting information)

8. Comparison between 2 groups of patients based on the observation of metastasis after surgery



Among a total of 238 patients without observed metastasis, 73 of them died.

Among a total of 160 patients with observed metastasis, 99 of them died with 38% chance of death at 44.7 months.

Logrank test result shows a significant difference in survival between these groups of patients (p=2e-08). *(see code in supporting information)*

To conclude, the 7 following variable Underlying_lung , AG, cT, cN, pN, pM, and meta are considered statistically significant.

Part 3: Model building and diagnosis

Objective:

1. Build and select the optimum multivariate model for lung cancer prognosis.
2. Evaluate the quality and the prediction power of the selected prognosis model.

Methods used:

1. Cox proportional hazards model: explore the relationships between the survival and explanatory variables.
2. Backward selection based on AIC (Akaike information criterion): select the optimum combination of covariate for multivariate model.
3. ROC (Receiver Operating Characteristics) – AUC (Area Under the Curve): define the degree of separability between class of event.
4. Case deletion residuals: verify influential observation.

Result and discussion:

To define the optimum combination of covariate for multivariate model, we started with the full model which includes all covariates for the prediction of survival time *(see code in supporting information)*. Then we did the automatic backward model selection based on AIC value. After 8 steps, the AIC converted from **1846.56** to **1827.36** with the combination of 11 covariates. Model summary showed the concordance score of this model as **0.732** with standard error **0.02** *(see code in supporting information)*. The details of features in multivariate model for cancer prognosis are shown in Table 4.

Table 4: Relationship between radiologic/clinicopathologic features and survival time (multivariate analysis)

Variables	Comparison group(s)	Reference group	P-value	Odds ratio
Age	**Age over than 60 years old**	**Age less than 60 years old**	**0.0067**	**1.827 [1.1819-2.824]**
Sex	Male	Female	0.0701	0.430 [0.1724-1.072]
Smoking state	**Current smoker**	**Non-smoker**	**0.0025**	**4.721 [1.7278-12.898]**
	Ex-smoker		**0.0109**	**3.481 [1.3321-9.098]**
	Unknown		**0.0063**	**3.739 [1.4503-9.642]**
Pulmonary function test (% FEV1/FVC)	numerical value		**0.0199**	**1.018 [1.0028-1.034]**
Metastasis	**Present**	**Absent**	**2.26e-07**	**2.3138 [1.6842-3.179]**
Criteria for size of original tumor in Central Lung SCC	**cT2**	**cT1**	**0.0206**	**1.8249 [1.0967-3.037]**
	cT3		**0.00030**	**2.944 [1.6398-5.285]**
	cT4		0.1000	1.985 [0.877-4.491]
Criteria for distant metastasis involved in Central Lung SCC	cM1b	cM0	0.1919	0.206 [0.0192-2.210]
Criteria for nearby lymph nodes involved in Peripheral Lung SCC	**pN1**	**pN0**	**0.0285**	**1.5441 [1.0467-2.278]**
	pN2		**1.20e-05**	**2.6081 [1.6979-4.006]**
	pN3		**0.0398**	**3.6911 [1.0627-12.821]**
Criteria for distant metastasis involved in Peripheral Lung SCC	**pM1b**	**pM0**	**0.0001**	**11.490 [3.2683-40.393]**
Present of obstructive pneumonitis/atelectasis	Present	Absent	0.0702	0.728 [0.5160-1.027]
Underlying lung disease	**Emphysema**	**No Remark**	**0.0245**	**1.551 [1.0582-2.275]**
	Interstitial lung abnormality (ILA)		**1.08e-07**	**4.744 [2.6715-8.426]**

Features that show statistically significant relationship with survival time in this multivariate Cox model are age, smoking state, pulmonary function test, metastasis, size of original tumor in central lung cancer, nearby lymph node involvement in peripheral lung cancer and underlying lung disease. Details are described as follow.

Patients older than 60 have approximately 2 times higher risk, which means a shorter survival time, compared to patients younger than 60 for every increase of 1 year (p-value=**0.0067**).
Current smoker and ex-smoker patients have approximately 5- and 3-times higher risk than non-smoker patients with p-value **0.0025** and **0.0109**, respectively.
Pulmonary function test indicates that every increasing unit of FEV1/FVC ratio unit will also increase the risk of approximately 1% (p-value=**0.0199**).
For both central and peripheral lung SCC, patients who have cancer spreading outside the lung have approximately twice higher risk than those without metastasis (p-value=**2.26e-07**).
Central lung cancer patients who have size of original tumor in criteria T2, T3 have 2- and 3-times higher risk than patients with criteria T1 with p-value **0.0206** and **0.0003**, respectively.
Nearby lymph nodes that are involved in peripheral lung cancer patients (pN) also play roles in survival prediction. Peripheral lung cancer patients whom are classified as criteria N1, N2 and N3 have approximately 1.5-, 2.6- and 3.7-times higher risk than patients who show no nodal involvement with p-value **0.0285**, **1.20e-05** and **0.0398**, respectively.
Peripheral lung cancer patients whom cancer spread to area outside the chest (pM1b) have approximately 11 times higher risk than patients with no cancer spreading outside the lung with p-value **0.0001**.
Patients with emphysema and the one with interstitial lung abnormality (ILA) show approximately 2- and 5- times higher risk than those without any underlying lung disease (p-value=**0.0245** and **1.08e-07**, respectively).

Besides the information about AIC and concordance index as discussed previously, we will do some additional tests to diagnose our model.
Result of ROC-AUC evaluation in Figure 1. ROC shows a probability curve while AUC which represents the degree of separability. It tells how much our model is capable of distinguishing between death status. Our model shows AUC 0.796 with 95% CI between **0.752-0.839** which means the model has approximately **80%** accuracy. Analysis of case deletion residual (Figure 2) observes no extreme observations as influential that can pull the regression toward them.

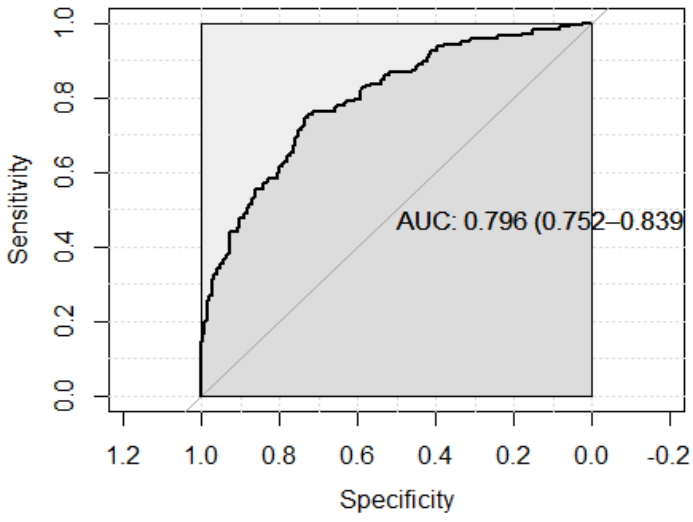


Figure 1: ROC-AUC evaluation of multiple cox regression model

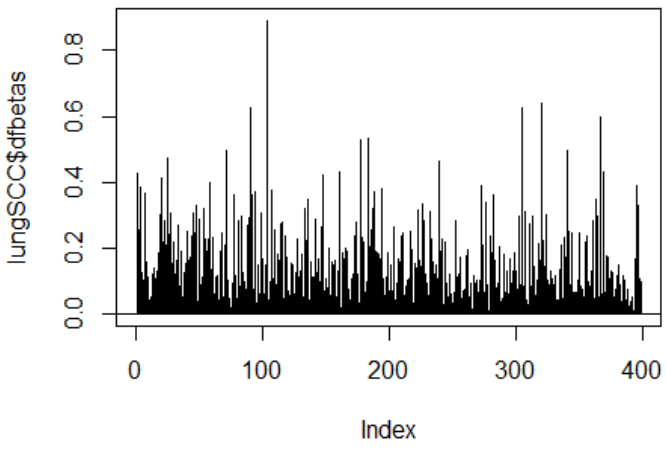


Figure 2: Analysis of case deletion residual

Conclusion:

Based on our analysis, age of patients, underlying lung disease with interstitial lung abnormality, metastasis condition, size of original tumor in central lung SCC patients, nearby lymph nodes involvement and distant metastasis in peripheral lung SCC can be classified as strong biomarker prediction in survival time of lung squamous cell carcinoma.

Supplement information:

```
library(tidyverse)
library(survival)
library(survminer)
library(asaur)
```

Data preparation: load the lung cancer dataset and check data types

```
lungSCC_raw <- read_csv("../LungCancer.csv",
                        col_types=cols(Sex='c',
                                       Age='c',
                                       Location='d',
                                       Obst_pn_or_plugging='d',
                                       Necrosis='d',
                                       Underlying_lung='d',
                                       Effusion='d',
                                       LN_ratio='d',
                                       cT='c',
                                       cN='c',
                                       cM='c',
                                       pT='c',
                                       pN='c',
                                       pM='c',
                                       Smoking_state='c',
                                       FEV1_FVC='d',
                                       CEA='d',
                                       op_type='c',
                                       death='d',
                                       OP_site_recur='l',
                                       meta='l',
                                       OS = "d",
                                       .default = 'c'))
```

Explore the dataset:

```
dim(lungSCC_raw)
# Radiologic features
table(lungSCC_raw$Location, useNA="always")
table(lungSCC_raw$Obst_pn_or_plugging, useNA="always")
table(lungSCC_raw$Necrosis, useNA="always")
table(lungSCC_raw$Underlying_lung, useNA="always")
table(lungSCC_raw$Effusion, useNA="always")
# Clinicopathologic features
table(lungSCC_raw$Sex, useNA="always")
table(lungSCC_raw$AG, useNA="always")
summary(lungSCC_raw$AG)
table(lungSCC_raw$LN_ratio, useNA="always")
summary(lungSCC_raw$LN_ratio)
table(lungSCC_raw$cT, useNA="always")
table(lungSCC_raw$cN, useNA="always")
table(lungSCC_raw$cM, useNA="always")
table(lungSCC_raw$pT, useNA="always")
table(lungSCC_raw$pN, useNA="always")
table(lungSCC_raw$pM, useNA="always")
table(lungSCC_raw$Smoking_state, useNA="always")
table(lungSCC_raw$FEV1_FVC, useNA="always")
summary(lungSCC_raw$FEV1_FVC)
table(lungSCC_raw$CEA)
summary(lungSCC_raw$CEA)
table(lungSCC_raw$OP, useNA="always")
table(lungSCC_raw$death, useNA="always")
table(lungSCC_raw$meta, useNA="always")
table(lungSCC_raw$OS, useNA="always")
summary(lungSCC_raw$OS)
```

Redefine the meaning of each label to make it easier for interpretation

```
lungSCC <- mutate(lungSCC_raw,
                  Location=ifelse(Location==0, "Peripheral", "Central"),

                  Sex=factor(Sex, levels=c("F", "M")),

                  Necrosis = factor(Necrosis, levels = c("0", "1", "2"), labels = c("NoRemark", "Necros
is", "Cavitation")),
                  Necrosis <- relevel(Necrosis, "NoRemark"),

                  Underlying_lung = factor(Underlying_lung, levels = c("0", "1", "2"), labels = c("NoRe
mark", "Emphysema", "ILA")),
                  Underlying_lung <- relevel(Underlying_lung, "NoRemark"),

                  Effusion=factor(Effusion, levels=c("0", "1"), labels=c("Absent", "Present")),

                  AG = ifelse((Age<60), "LT60", "OV60"),
                  AG = factor(AG),

                  OP = factor(op_type, levels = c("5", "1", "2", "3", "4"), labels = c("No op", "Lobect
omy", "Pneumonectomy", "Segmentectomy", "LN biopsy")),
                  OP <- relevel(OP, "No op"),

                  Smoking_state = factor(Smoking_state, levels = c("3", "0", "1", "2"), labels = c("Non
-smoker", "Current-smoker", "Ex-smoker", "Unknown")),
                  Smoking_state <- relevel(Smoking_state, "Non-smoker"),

                  cT = factor(cT, levels = c("1", "2", "3", "4"), labels = c("cT1", "cT2", "cT3", "cT4"
)),

                  cN = factor(cN, levels = c("1", "2", "3", "4"), labels = c("cN0", "cN1", "cN2", "cN3"
)),

                  cM = factor(cM, levels = c("1", "3"), labels = c("cM0", "cM1b")),

                  pT = factor(pT, levels = c("1", "2", "3", "4"), labels = c("pT1", "pT2", "pT3", "pT4"
)),

                  pN = factor(pN, levels = c("1", "2", "3", "4"), labels = c("pN0", "pN1", "pN2", "pN3"
)),

                  pM = factor(pM, levels = c("1", "3"), labels = c("pM0", "pM1b")))
```

Part 1: Relationship between radiologic/clinicopathologic features and death status

Example of code for Fisher's Exact Test

Difference in risk of death in different age group

```
with(lungSCC, fisher.test(table(death, AG)))
```

Example of code for Fisher's Exact Test

Difference in risk of death in different age group

```
fit_underlying_lung <- glm(death~Underlying_lung, family="binomial", data=lungSCC)
summary(fit_underlying_lung)
```

Odds ratio

```
exp(0.5125)
exp(2.2807)
```

Part 2: Relationship between radiologic/clinicopathologic features and survival time (univariate analysis)

Part2. A. Cox Proportional Hazards Regression Model fit

Example: location vs OS: tumor location:

```
fit_location <- coxph(Surv(OS, death) ~ Location, data=lungSCC)
summary(fit_location)
```

Part2. B. Kaplan-Meier estimator and Logrank test with features which have a significant relationship with survival time

Example: comparison between 2 groups of patients based on underlying lung disease:

```
fit.KM_under <- survfit(Surv(OS, death) ~ Underlying_lung, data=lungSCC)
fit.KM_under
```

We plot the Kaplan-Meier survival curve:

```
gg <- ggsurvplot(fit.KM_under,
           legend = "bottom",
           conf.int = TRUE,
           linetype = "strata",
           legend.title = "Underlying lung disease",
           xlab = "Months",
           ylab = "Survival probability")
gg

ggsave("KM_Underlying lung disease.png", plot=print(gg))
```

The logrank test:

```
survdiff(Surv(OS, death) ~ Underlying_lung, data=lungSCC)
```

Part 3: Model building and diagnosis

Code for backward selection to find the optimum multivariate Cox model

Start with the full model with all variables

```
lungSCC_fullmodel <- coxph(Surv(OS, death) ~ Location + Obst_pn_or_plugging + Necrosis + Underlying_lun
g + Effusion + Sex + AG + LN_ratio + cT+ cN + cM + pT+ pN+ pM + Smoking_state + FEV1_FVC + CEA + meta ,
 data = lungSCC)
```

Automatic model selection based on AIC

```
lungSCC_multiCox <- step(lungSCC_fullmodel)
```

Code for summary multivariate model

```
summary(lungSCC_multiCox)
```

Check prediction power of the model with ROC-AUC

```
library(pROC)

lungSCC_predict <- select(lungSCC, -death, -OS)
lungSCC_predict$lp <- predict(lungSCC_multiCox, newdata=lungSCC_predict, type="lp")


roc(lungSCC$death,lungSCC_predict$lp,
    smoothed = TRUE,
    ci = TRUE, ci.alpha = 0.9, stratified = FALSE,
    plot = TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid = TRUE,
    print.auc=TRUE, show.thres=TRUE)
```

Model diagnosis with case deletion residual

```
dfbetas <- residuals(lungSCC_multiCox, type = 'dfbetas')
lungSCC$dfbetas <- sqrt(rowSums(dfbetas^2))
plot(lungSCC$dfbetas, type = 'h')
abline(h = 0)
```